Sylvia Wassertheil-Smoller
Jordan Smoller

Biostatistics and Epidemiology

A Primer for Health and Biomedical Professionals

Fifth Edition

Springer

Biostatistics and Epidemiology

Sylvia Wassertheil-Smoller • Jordan Smoller

Biostatistics and Epidemiology

A Primer for Health and Biomedical Professionals

Fifth Edition

Springer

Sylvia Wassertheil-Smoller
Department of Epidemiology
and Population Health
Albert Einstein College of Medicine
Bronx, NY, USA

Jordan Smoller
Department of Psychiatry and Center
for Genomic Medicine
Massachusetts General Hospital
Boston, MA, USA

ISBN 978-3-031-53042-5 ISBN 978-3-031-53043-2 (eBook)
https://doi.org/10.1007/978-3-031-53043-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

To Alexis and Ava

Preface to the Fifth Edition

This book, through its several editions, has evolved more or less in parallel to the evolving statistical and epidemiologic methodologies, often driven by huge advances in computing power. Thus, the book includes both very basic concepts, as well as introduction to complex, newer methods. Our continuing aim is to adapt to these evolving areas of research in epidemiology and statistics, while maintaining the original objective of being non-threatening, understandable, and accessible to those with limited or no background in mathematics. Some new areas covered in this fifth edition include new material on penalized regression used in machine learning, and new material on mediation and moderation, and expanded and updated topics in genetic epidemiology.

Causal inference is an important objective of epidemiology and biomedical research. The randomized clinical trial is the "gold standard" for evidence on causation and on comparing treatments. However, we are not able to do clinical trials in all areas, either due to feasibility issues, high costs, or sample size and length of follow-up time required to draw valid conclusions. Thus, we must often rely on evidence from observational studies that may be subject to confounding. Propensity analysis is one analytical technique used to control for confounding. Increasingly, Mendelian randomization is being used to infer causation in observational studies and this fifth edition provides an explanation of the methods involved. Another development in epidemiology has been the increasing availability of very large datasets—for example, those that examine electronic health records (EHR). To analyze such data, researchers often turn to machine learning methods that can extract signals from a very large number of variables. This edition includes new sections that describe some of these methods such as penalized regression techniques like LASSO or Ridge regressions to shrink the variable set.

The principal objectives of the earlier editions still apply. The presentation of the material is aimed to give an understanding of the underlying principles, as well as practical guidelines of "how to do it" and "how to interpret it." The topics included are those that are most commonly used or referred to in the literature. There are some features to note that may aid the reader in the use of this book:

(a) The book starts with a discussion of the philosophy and logic of science and the underlying principles of testing what we believe against the reality of our experiences. While such a discussion, per se, will not help the reader to actually "do a *t*-test," we think it is important to provide some introduction to the underlying framework of the field of epidemiology and statistics, to understand why we do, what we do.
(b) Many of the subsections stand alone; that is, the reader can turn to the topic that interests him or her and read the material out of sequential order. Thus, the book may be used by those who need it for special purposes. The reader is free to skip those topics that are not of interest without being too much hampered in further reading. By design, then, there is some redundancy. By our teaching experience, however, we have found that it is better to err on the side of redundancy than on the side of sparsity.
(c) Cross-references to other relevant sections are included when additional explanation is needed. When development of a topic is beyond the scope of this text, the reader is referred to other books that deal with the material in more depth or on a higher mathematical level.
(d) The appendices provide sample calculations for various statistics described in the text. This makes for smoother reading of the text, while providing the reader with more specific instructions on how actually to do some of the calculations.

The prior editions grew from feedback from students who indicated they appreciated the clarity and the focus on topics specifically related to their work. However, some users missed coverage of several important topics. Accordingly, sections were added to include a full chapter on measures of quality of life and various psychological scales, which are increasingly used in clinical studies; an expansion of the chapter on probability, with the introduction of several nonparametric methods; the clarification of some concepts that were more tersely addressed previously; and the addition of several appendices (providing sample calculations of the Fisher's exact test, Kruskal–Wallis test, and various indices of reliability and responsiveness of scales used in quality of life measures).

It requires a delicate balance to keep the book concise and basic, and yet make it sufficiently inclusive to be useful to a wide audience. We hope this book will be useful to diverse groups of people in the health field, as well as to those in related areas. The material is intended for: (1) physicians doing clinical research, as well as for those doing basic research; (2) students—medical, college, and graduate; (3) research staff in various capacities; (4) those interested in the growing field of

genetic epidemiology and wanting to be able to read genetic research or wishing to collaborate in genetic research; and (5) anyone interested in the logic and methodology of biostatistics, epidemiology, and genetic epidemiology. The principles and methods described here are applicable to various substantive areas, including medicine, public health, psychology, and education.

Bronx, NY, USA — Sylvia Wassertheil-Smoller
Boston, MA, USA — Jordan Smoller

Acknowledgments

I want to express my gratitude for the inspired teaching of Dr. Jacob Cohen, now deceased, who started me on this path, and to my colleagues and students at the Albert Einstein College of Medicine, who make it fun.

My appreciation goes to those colleagues who critiqued the earlier editions, with special thanks to Dr. Aileen McGinn, Dr. Gloria Ho, Dr. Tao Wang, and Dr. Kenny Ye as well as to Dr. Xiaon Xue and Dr. Jianwen Cai for their clarity, willingness to answer questions, and their wise suggestions for this fifth edition.

Sadly, my late husband, Walter Austerer, is not here to enjoy this new edition, but I wish to honor his memory and the patience, love, and support he unfailingly gave through previous editions.

Finally, I want to say what a great privilege and pleasure it is to work with my co-author who is my son, Jordan Smoller. Life rarely brings such rewards.

Bronx, NY, USA Sylvia Wassertheil-Smoller

Contents

About the Authors

Sylvia Wassertheil-Smoller Ph.D. is Professor of Epidemiology in the Department of Epidemiology, and Population Health at the Albert Einstein College of Medicine and is the Dorothy and William Manealoff Foundation and Molly Rosen Chair in Social Medicine, Emerita. In addition to her teaching, her research areas span both cardiovascular disease and cancer epidemiology. She has been a leader in landmark clinical trials that informed guidelines for the prevention of heart disease and stroke, and in major national and international collaborative prospective studies. She has won awards for mentoring students and junior faculty, as well as the Einstein Spirit of Achievement Award. She lives in New York.

Jordan Smoller M.D., Sc.D. is Professor of Psychiatry at Harvard Medical School and Professor in the Department of Epidemiology at the Harvard T.H. Chan School of Public Health in Boston. He is Director of the Psychiatric and Neurodevelopmental Genetics Unit in the Massachusetts General Hospital (MGH) Center for Genomic Medicine, Director of the Center for Precision Psychiatry at MGH, and an Associate Member of the Broad Institute. The focus of Dr. Smoller's research is understanding the genetic and environmental determinants of psychiatric disorders across the lifespan and using big data to advance precision mental health, including improved methods to reduce risk and enhance resilience. He is the recipient of numerous research awards; an author of more than 500 scientific articles, book chapters, and reviews; as well as the author of the book *The Other Side of Normal* (HarperCollins/William Morrow, 2012). He lives with his family in Boston.

Chapter 1
The Scientific Method

Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.

Jules Henri Poincare
La Science et l'Hypothese (1908)

1.1 The Logic of Scientific Reasoning

The whole point of science is to uncover the "truth." How do we go about deciding something is true? We have two tools at our disposal to pursue scientific inquiry:

We have our senses, through which we experience the world and make *observations*.
We have the ability to reason, which enables us to make logical *inferences*.
In science we impose *logic* on those observations.

Clearly, we need both tools. All the logic in the world is not going to create an observation, and all the individual observations in the world won't in themselves create a theory. There are two kinds of relationships between the scientific mind and the world and two kinds of logic we impose—*deductive and inductive*—as illustrated in Figure 1.1.

In *deductive inference*, we hold a theory, and based on it, we make a prediction of its consequences. That is, we predict what the observations should be. For example, we may hold a theory of learning that says that positive reinforcement results in better learning than does punishment; that is, rewards work better than punishments. From this theory, we predict that math students who are praised for their right answers during the year will do better on the final exam than those who are punished for their wrong answers. We go from the general, the theory, to the specific, the observations. This is known as the hypothetico-deductive method.

In *inductive inference*, we go from the specific to the general. We make many observations, discern a pattern, make a generalization, and infer an explanation. For example, it was observed in the Vienna General Hospital in the 1840s that women giving birth were dying at a high rate of puerperal fever, a generalization that provoked terror in prospective mothers. It was a young doctor named Ignaz Phillip

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_1

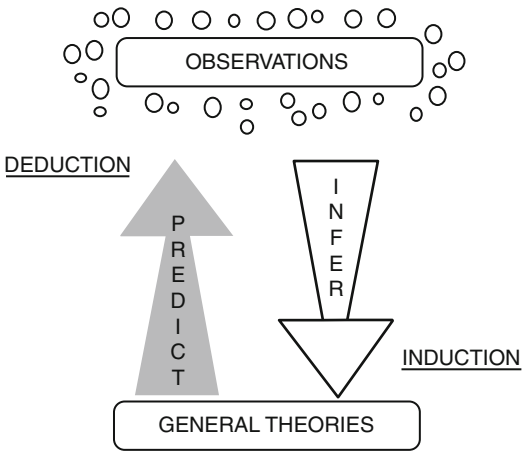


Figure 1.1 Deductive and inductive inference

Semmelweis who connected the observation that medical students performing vaginal examinations did so directly after coming from the dissecting room, rarely washing their hands in between, with the observation that a colleague who accidentally cut his finger while dissecting a corpse died of a malady exactly like the one killing the mothers. He inferred the explanation that the cause of death was the introduction of cadaverous material into a wound. The practical consequence of that creative leap of the imagination was the elimination of puerperal fever as a scourge of childbirth by requiring that physicians wash their hands before doing a delivery! The ability to make such creative leaps from generalizations is the product of creative scientific minds.

Epidemiologists have generally been thought to use inductive inference. For example, several decades ago, it was noted that women seemed to get heart attacks about 10 years later than men did. A creative leap of the imagination led to the inference that it was women's hormones that protected them until menopause. EUREKA! They deduced that if estrogen was good for women, it must be good for men and predicted that the observations would corroborate that deduction. A clinical trial was undertaken, which gave men at high risk of heart attack estrogen in rather large doses, 2.5 mg per day or about four times the dosage currently used in postmenopausal women. Unsurprisingly, the men did not appreciate the side effects, but surprisingly to the investigators, the men in the estrogen group had higher coronary heart disease rates and mortality than those on placebo.[2] What was good for the goose might not be so good for the gander. The trial was discontinued, and estrogen as a preventive measure was abandoned for several decades.

During that course of time, many prospective observational studies indicated that estrogen replacement given to postmenopausal women reduced the risk of heart disease by 30–50%. These observations led to the inductive inference that postmenopausal hormone replacement is protective, i.e., observations led to theory. However, that theory must be tested in clinical trials. The first such trial of hormone replacement in women who already had heart disease, the Heart and Estrogen/

progestin Replacement Study (HERS), found no difference in heart disease rates between the active treatment group and the placebo group, but did find an early increase in heart disease events in the first year of the study and a later benefit of hormones after about 2 years. Since this was a study in women with established heart disease, it was a secondary prevention trial and does not answer the question of whether women without known heart disease would benefit from long-term hormone replacement. That question has been addressed by the Women's Health Initiative (WHI), which is described in a later section.

The point of the example is to illustrate how observations (that women get heart disease later than men) lead to theory (that hormones are protective), which predicts new observations (that there will be fewer heart attacks and deaths among those on hormones), which may strengthen the theory, until it is tested in a clinical trial that can either corroborate it or overthrow it and lead to a new theory, which then must be further tested to see if it better predicts new observations. So there is a constant interplay between inductive inference (based on observations) and deductive inference (based on theory), until we get closer and closer to the "truth."

However, there is another point to this story. Theories don't just leap out of facts. There must be some substrate out of which the theory leaps. Perhaps that substrate is another preceding theory that was found to be inadequate to explain these new observations and that theory, in turn, had replaced some previous theory. In any case, one aspect of the "substrate" is the "prepared mind" of the investigator. If the investigator is a cardiologist, for instance, he or she is trained to look at medical phenomena from a cardiology perspective and is knowledgeable about preceding theories and their strengths and flaws. If the cardiologist hadn't had such training, he or she might not have seen the connection. Or, with different training, the investigator might leap to a different inference altogether. The epidemiologist must work in an interdisciplinary team to bring to bear various perspectives on a problem and to enlist minds "prepared" in different ways.

The question is, how well does a theory hold up in the face of new observations? When many studies provide affirmative evidence in favor of a theory, does that increase our belief in it? Affirmative evidence means more examples that are consistent with the theory. But to what degree does supportive evidence strengthen an assertion? Those who believe induction is the appropriate logic of science hold the view that affirmative evidence is what strengthens a theory.

Another approach is that of Karl Popper,[1] perhaps one of the foremost theoreticians of science. Popper claims that induction arising from accumulation of affirmative evidence doesn't strengthen a theory. Induction, after all, is based on our belief that the things unobserved will be like those observed or that the future will be like the past. For example, we see a lot of white swans and we make the assertion that all swans are white. This assertion is supported by many observations. Each time we see another white swan, we have more supportive evidence. But we cannot prove that all swans are white no matter how many white swans we see.

On the other hand, this assertion can be knocked down by the sighting of a single black swan. Now we would have to change our assertion to say that most swans are white and that there are some black swans. This assertion presumably is closer to the

truth. In other words, we can refute the assertion with one example, but we can't prove it with many. (The assertion that all swans are white is a descriptive generalization rather than a theory. A theory has a richer meaning that incorporates causal explanations and underlying mechanisms. Assertions, like those relating to the color of swans, may be components of a theory.)

According to Popper, the proper methodology is to posit a theory, or a conjecture, as he calls it, and try to demonstrate that it is false. The more such attempts at destruction it survives, the stronger is the evidence for it. The object is to devise ever more aggressive attempts to knock down the assertion and see if it still survives. If it does not survive an attempt at *falsification*, then the theory is discarded and replaced by another. He calls this the method of *conjectures and refutations*. The advance of science toward the "truth" comes about by discarding theories whose predictions are not confirmed by observations or theories that are not testable altogether, rather than by shoring up theories with more examples of where they work. *Useful scientific theories are potentially falsifiable.*

Untestable theories are those where a variety of contradictory observations could each be consistent with the theory. For example, consider Freud's psychoanalytic theory. The Oedipus complex theory says that a child is in love with the parent of the opposite sex. A boy desires his mother and wants to destroy his father. If we observe a man to say he loves his mother, that fits in with the theory. If we observe a man to say he hates his mother, that also fits in with the theory, which would say that it is "reaction formation" that leads him to deny his true feelings. In other words, no matter what the man says, it could not falsify the theory because it could be explained by it. Since no observation could potentially falsify the Oedipus theory, its position as a scientific theory could be questioned.

A third, and most reasonable, view is that the progress of science requires both inductive and deductive inference. A particular point of view provides a framework for observations, which lead to a theory that predicts new observations that modify the theory, which then leads to new, predicted observations, and so on toward the elusive "truth," which we generally never reach. Asking which comes first, theory or observation, is like asking which comes first, the chicken or the egg.

In general then, advances in knowledge in the health field come about through constructing, testing, and modifying theories. Epidemiologists make inductive inferences to generalize from many observations, make creative leaps of the imagination to infer explanations and construct theories, and use deductive inferences to test those theories.

Theories, then, can be used to predict observations. But these observations will not always be exactly as we predict them, due to error and the inherent variability of natural phenomena. If the observations are widely different from our predictions, we will have to abandon or modify the theory. How do we test the extent of the discordance of our predictions based on theory from the reality of our observations? The test is a statistical or probabilistic test. It is the test of *the null hypothesis*, *which is the cornerstone of statistical inference* and will be discussed later. Some excellent classic writings on the logic and philosophy of science, and applications in epidemiology, are listed in the "References" section at the end of this book, and while some were written quite a while ago, they are still obtainable.[2–7]

1.2 Variability of Phenomena Requires Statistical Analysis

Statistics is a methodology with broad areas of application in science and industry as well as in medicine and in many other fields. A phenomenon may be principally based on a deterministic model. One example is Boyle's law, which states that for a fixed volume an increase in temperature of a gas determines that there is an increase in pressure. Each time this law is tested, the same result occurs. The only variability lies in the error of measurement. Many phenomena in physics and chemistry are of such a nature.

Another type of model is a probabilistic model, which implies that various states of a phenomenon occur with certain probabilities. For instance, the distribution of intelligence is principally probabilistic, that is, given values of intelligence occur with a certain probability in the general population. In biology, psychology, or medicine, where phenomena are influenced by many factors that in themselves are variable and by other factors that are unidentifiable, the models are often probabilistic. In fact, as knowledge in physics has become more refined, it begins to appear that models formerly thought to be deterministic are probabilistic.

In any case, where the model is principally probabilistic, statistical techniques are needed to increase scientific knowledge. *The presence of variation requires the use of statistical analysis.*[7] When there is little variation with respect to a phenomenon, much more weight is given to a small amount of evidence than when there is a great deal of variation. For example, we know that pancreatic cancer appears to be invariably a fatal disease. Thus, if we found a drug that indisputably cured a few patients of pancreatic cancer, we would give a lot of weight to the evidence that the drug represented a cure, far more weight than if the course of this disease were more variable. In contrast to this example, if we were trying to determine whether vitamin C cures colds, we would need to demonstrate its effect in many patients, and we would need to use statistical methods to do so, since human beings are quite variable with respect to colds. In fact, in most biological and even more so in social and psychological phenomena, there is a great deal of variability.

1.3 Inductive Inference: Statistics as the Technology of the Scientific Method

Statistical methods are objective methods by which *group trends are abstracted from observations on many separate individuals*. A simple concept of statistics is the calculation of averages, percentages, and so on and the presentation of data in tables and charts. Such techniques for summarizing data are very important indeed and essential to describing the population under study. However, they make up a small part of the field of statistics. A major part of statistics involves the *drawing of inferences from samples to a population* in regard to some characteristic of interest. Suppose we are interested in the average blood pressure of women college students.

If we could measure the blood pressure of every single member of this population, we would not have to infer anything. We would simply average all the numbers we obtained. In practice, however, we take a sample of students (properly selected), and on the basis of the data we obtain from the sample, we infer what the mean of the whole population is likely to be.

The reliability of such inferences or conclusions may be evaluated in terms of probability statements. *In statistical reasoning, then, we make inductive inferences, from the particular (sample) to the general (population).* Thus, statistics may be said to be the technology of the scientific method.

1.4 Design of Studies

While the generation of hypotheses may come from anecdotal observations, the testing of those hypotheses must be done by making controlled observations, free of systematic bias. Statistical techniques, to be valid, must be applied to data obtained from well-designed studies. Otherwise, solid knowledge is not advanced.

There are two types of studies: The first is observational studies, where "Nature" determines who is exposed to the factor of interest and who is not exposed. These studies demonstrate association. Association may imply causation or it may not. The second is experimental studies, where the investigator determines who is exposed. These may prove causation.

Observational studies may be of three different study designs: *cross-sectional*, *case–control*, or *prospective*. In a *cross-sectional study*, the measurements are taken at one point in time. For example, in a cross-sectional study of high blood pressure and coronary heart disease, the investigators determine the blood pressure and the presence of heart disease at the same time. If they find an association, they would not be able to tell which came first. Does heart disease result in high blood pressure or does high blood pressure cause heart disease, or are both high blood pressure and heart disease the result of some other common cause?

In a *case–control study* of smoking and lung cancer, for example, the investigator starts with lung cancer cases and with controls, and through examination of the records or through interviews determines the presence or the absence of the factor in which he or she is interested (smoking). A case–control study is sometimes referred to as a *retrospective study* because data on the factor of interest are collected retrospectively and thus may be subject to various inaccuracies.

In a *prospective* (or *cohort*) study, the investigator starts with a cohort of nondiseased persons with that factor (i.e., those who smoke) and persons without that factor (nonsmokers) and goes forward into some future time to determine the frequency of development of the disease in the two groups. A prospective study is also known as a longitudinal study. *The distinction between case–control studies and prospective studies lies in the sampling. In the case–control study, we sample from among the diseased and nondiseased, whereas in a prospective study, we sample from among those with the factor and those without the factor*. Prospective

studies provide stronger evidence of causality than retrospective studies but are often more difficult, more costly, and sometimes impossible to conduct, for example, if the disease under study takes decades to develop or if it is very rare.

In the health field, an experimental study to test an intervention of some sort is called a *clinical trial*. In a clinical trial, the investigator assigns patients or participants to one group or another, usually randomly, while trying to keep all other factors constant or controlled for, and compares the outcome of interest in the two (or more) groups. More about clinical trials is in Chapter 6.

In summary, then, the following list is in ascending order of strength in terms of demonstrating causality:

- *Cross-sectional studies:* useful in showing associations, in providing early clues to etiology
- *Case–control studies:* useful for rare diseases or conditions or when the disease takes a very long time to become manifest (other name: *retrospective studies*)
- *Cohort studies:* useful for providing stronger evidence of causality and less subject to biases due to errors of recall or measurement (other names: *prospective studies*, *longitudinal studies*).
- *Clinical trials:* prospective, experimental studies that provide the most rigorous evidence of causality.

1.5 How to Quantify Variables

How do we test a hypothesis? First of all, we must set up the hypothesis in a *quantitative* manner. Our criterion measure must be a number of some sort. For example, how many patients died in a drug group compared with how many of the patients died who did not receive the drug, or what is the mean blood pressure of patients on a certain antihypertensive drug compared with the mean blood pressure of patients not on this drug. Sometimes variables are difficult to quantify. For instance, if you are evaluating the quality of care in a clinic in one hospital compared with the clinic of another hospital, it may sometimes be difficult to find a quantitative measure that is representative of quality of care, but nevertheless it can be done and it must be done if one is to test the hypothesis.

There are two types of data that we can deal with: *discrete* or *categorical variables* and *continuous variables*. Continuous variables, theoretically, can assume an infinite number of values between any two fixed points. For example, weight is a continuous variable, as is blood pressure, time, intelligence, and, in general, variables in which measurements can be taken. Discrete variables (or categorical variables) are variables that can only assume certain fixed numerical values. For instance, sex is a discrete variable. You may code it as 1 = male and 2 = female, but an individual cannot have a code of 1.5 on sex (at least not theoretically). Discrete variables generally refer to counting, such as the number of patients in a given group who live, the number of people with a certain disease, and so on. In

Chapter 3 we will consider a technique for testing a hypothesis where the variable is a discrete one, and, subsequently, we will discuss some aspects of continuous variables, but first we will discuss the general concepts of hypothesis testing.

1.6 The Null Hypothesis

The hypothesis we test statistically is called the *null hypothesis*. Let us take a conceptually simple example. Suppose we are testing the efficacy of a new drug on patients with myocardial infarction (heart attack). We divide the patients into two groups—drug and no drug—according to good design procedures and use as our criterion measure mortality in the two groups. It is our hope that the drug lowers mortality, but to test the hypothesis statistically, we have to set it up in a sort of backward way. We say our hypothesis is that the drug makes no difference, and what we hope to do is to reject the "no difference" hypothesis, based on evidence from our sample of patients. This is known as the *null hypothesis*. We specify our test hypothesis as follows:

H_o (null hypothesis) Death rate in group treated with drug A = death rate in group treated with drug B

This is equivalent to

H_o: (death rate in group A)—(death rate in group B) = 0

We test this against an *alternate hypothesis*, known as H_A, that the difference in death rates between the two groups *does not* equal 0.

We then gather data and note the *observed* difference in mortality between group A and group B. If this observed difference is sufficiently greater than zero, we reject the null hypothesis. If we reject the null hypothesis of no difference, we accept the *alternate hypothesis*, which is that the drug does make a difference.

When you test a hypothesis, this is the type of reasoning you use:

(1) I will *assume* the hypothesis that there is no difference is true.
(2) I will then collect the data and *observe* the difference between the two groups.
(3) If the null hypothesis is true, how likely is it that *by chance alone* I would get results such as these?
(4) If it is not likely that these results could arise by chance under the assumption than the null hypothesis is true, then I will conclude it is false, and I will "accept" the alternate hypothesis.

1.7 Why Do We Test the Null Hypothesis?

Suppose we believe that drug A is better than drug B in preventing death from a heart attack. Why don't we test that belief directly and see which drug is better rather than testing the hypothesis that drug A is *equal* to drug B? The reason is that there is an infinite number of ways in which drug A can be better than drug B, so we would have to test an infinite number of hypotheses. If drug A causes 10% fewer deaths than drug B, it is better. So first we would have to see if drug A causes 10% fewer deaths. If it doesn't cause 10% fewer deaths, but if it causes 9% fewer deaths, it is also better. Then we would have to test whether our observations are consistent with a 9% difference in mortality between the two drugs. Then we would have to test whether there is an 8% difference and so on. Note: Each such hypothesis would be set up as a null hypothesis in the following form: drug A—drug B mortality = 10%, or equivalently

$$\begin{aligned}
&(\text{drug A} - \text{drug B mortality}) - (10\%) = 0.\\
&(\text{drug A} - \text{drug B mortality}) - (9\%) = 0.\\
&(\text{drug A} - \text{drug B mortality}) - (8\%) = 0.
\end{aligned}$$

On the other hand, when we test the null hypothesis of no difference, we only have to test one value—a 0% difference—and we ask whether our observations are consistent with the hypothesis that there is *no* difference in mortality between the two drugs. If the observations are consistent with a null difference, then we cannot state that one drug is better than the other. If it is unlikely that they are consistent with a null difference, then we can reject that hypothesis and conclude there is a difference.

A common source of confusion arises when the investigator really wishes to show that one treatment is as good as another (in contrast to the above example, where the investigator in her heart of hearts really believes that one drug is better). For example, in the emergency room, a quicker procedure may have been devised and the investigator believes it may be as good as the standard procedure, which takes a long time. The temptation in such a situation is to "prove the null hypothesis." *But it is impossible to "prove" the null hypothesis.*

All statistical tests can do is reject the null hypothesis or fail to reject it. We do not prove the hypothesis by gathering affirmative or supportive evidence, because no matter how many times we did the experiment and found a difference close to zero, we could never be assured that the next time we did such an experiment, we would not find a huge difference that was nowhere near zero. It is like the example of the white swans discussed earlier: No matter how many white swans we see, we cannot prove that all swans are white, because the next sighting might be a black swan. Rather, we try to falsify or reject our assertion of no difference, and if the assertion of zero difference withstands our attempt at refutation, it survives as a hypothesis in which we continue to have belief. Failure to reject it does not mean we have proven that there is really no difference. It simply means that the evidence we have "is

consistent with" the null hypothesis. The results we obtained could have arisen by chance alone if the null hypothesis were true. (Perhaps the design of our study was not appropriate. Perhaps we did not have enough patients.)

So what can one do if one really wants to show that two treatments are equivalent? *One can design a study that is large enough to detect a small difference if there really is one.* If the study has the power (meaning a high likelihood) to detect a difference that is very, very, very small, and one fails to detect it, then one can say with a high degree of confidence that one can't find a meaningful difference between the two treatments. It is impossible to have a study with sufficient power to detect a 0% difference. As the difference one wishes to detect approaches zero, the number of subjects necessary for a given power approaches infinity. The relationships among significance level, power, and sample size are discussed more fully in Chapter 6.

1.8 Types of Errors

The important point is that *we can never be certain* that we are right in either accepting or rejecting a hypothesis. In fact, we run the risk of making one of two kinds of errors: We can reject the null or test hypothesis incorrectly; that is, we can conclude that the drug does reduce mortality when in fact it has no effect. This is known as a *type I error*. Or we can fail to reject the null or test hypothesis incorrectly; that is, we can conclude that the drug does not have an effect when in fact it does reduce mortality. This is known as a *type II error*. Each of these errors carries with it certain consequences. In some cases, a type I error may be more serious; in other cases, a type II error may be more serious. These points are illustrated in Figure 1.2.

Null Hypothesis (H_o)
Drug has no effect—no difference in mortality between patients using drug and patients not using drug.

		TRUE STATE OF NATURE	
		DRUG HAS NO EFFECT H_o True	DRUG HAS EFFECT; H_o False, H_A True
DECISION ON BASIS OF SAMPLE	DO NOT REJECT H_o No Effect	NO ERROR	TYPE II ERROR
	REJECT H_o (Accept H_A) Effect	TYPE I ERROR	NO ERROR

Figure 1.2 Hypothesis testing and types of error

Alternate Hypothesis (H_A)
Drug has effect—reduces mortality.

If we don't reject H_o, we conclude there is no relationship between drug and mortality. If we do reject H_O and accept H_A, we conclude there is a relationship between drug and mortality.

Actions to Be Taken Based on Decision

(1) If we believe the null hypothesis (i.e., fail to reject it), we will not use the drug.
*Consequences of **wrong** decision:* type II error. If we believe H_o incorrectly, since in reality the drug is beneficial, by withholding it we will allow patients to die who might otherwise have lived.

(2) If we reject null hypothesis in favor of the alternate hypothesis, we will use the drug.
*Consequences of **wrong** decision:* type I error. If we have rejected the null hypothesis incorrectly, we will use the drug and patients don't benefit. Presuming the drug is not harmful in itself, we do not directly hurt the patients, but since we think we have found the cure, we might no longer test other drugs.

We can never absolutely know the "True State of Nature," but we infer it on the basis of sample evidence.

1.9 Significance Level and Types of Error

We cannot eliminate the risk of making one of these kinds of errors, but we can lower the probabilities that we will make these errors. *The probability of making a type I error is known as the significance level of a statistical test.* When you read in the literature that a result was significant at the 0.05 level, it means that in this experiment, the results are such that the probability of making a type I error is less than or equal to 0.05. Mostly in experiments and surveys, people are very concerned about having a low probability of making a type I error and often ignore the type II error. This may be a mistake since in some cases a type II error is a more serious one than a type I error. In designing a study, if you aim to lower the type I error, you automatically raise the type II error probability. To lower the probabilities of both the type I and type II error in a study, it is necessary to increase the number of observations.

It is interesting to note that the rules of the Food and Drug Administration (FDA) are set up to lower the probability of making type I errors. In order for a drug to be approved for marketing, the drug company must be able to demonstrate that it does no harm and that it is effective. Thus, many drugs are rejected because their effectiveness cannot be adequately demonstrated. The null hypothesis under test is "this drug makes no difference." To satisfy FDA rules, this hypothesis must be rejected, with the probability of making a type I error (i.e., rejecting it incorrectly) being quite low. In other words, the FDA doesn't want a lot of useless drugs on the market. Drug companies, however, also give weight to guarding against type II error (i.e., avoid believing the no-difference hypothesis incorrectly) so that they may market potentially beneficial drugs.

1.10 Consequences of Type I and Type II Errors

The relative seriousness of these errors depends on the situation. Remember, a type I error (also known as *alpha*) means you are stating something is really there (an effect) when it actually is not, and a type II error (also known as *beta* error) means you are missing something that is really there. If you are looking for a cure for cancer, a type II error would be quite serious. You would miss finding useful treatments. If you are considering an expensive drug to treat a cold, clearly you would want to avoid a type I error; that is, you would not want to make false claims for a cold remedy.

It is difficult to remember the distinction between type I and II errors. Perhaps this small parable will help us. Once there was a King who was very jealous of his Queen. He had two knights, Alpha, who was very handsome, and Beta, who was very ugly. It happened that the Queen was in love with Beta. The King, however, suspected the Queen was having an affair with Alpha and had him beheaded. Thus, the King made both kinds of errors: He suspected a relationship (with Alpha) where there was none, and he failed to detect a relationship (with Beta) where there really was one. The Queen fled the kingdom with Beta and lived happily ever after, while the King suffered torments of guilt about his mistaken and fatal rejection of Alpha.

More on alpha, beta, power, and sample size appear in Chapter 6. Since hypothesis testing is based on probabilities, we will first present some basic concepts of probability in Chapter 2.

Chapter 2
A Little Bit of Probability

The theory of probability is at bottom nothing but common sense reduced to calculus.

Pierre Simon De Le Place
Theori Analytique des Probabilites (1812–1820)

2.1 What Is Probability?

The probability of the occurrence of an event is indicated by a number ranging from 0 to 1. An event whose probability of occurrence is 0 is certain not to occur, whereas an event whose probability is 1 is certain to occur.

The classical definition of probability is as follows: If an event can occur in N mutually exclusive, equally likely ways and if n_A of these outcomes have attribute A, then the probability of A, written as $P(A)$, equals n_A/N. This is an a priori definition of probability; that is, one determines the probability of an event before it has happened. Assume one were to toss a die and wanted to know the probability of obtaining a number divisible by three on the toss of a die. There are six possible ways that the die can land. Of these, there are two ways in which the number on the face of the die is divisible by three, a 3 and a 6. Thus, the probability of obtaining a number divisible by three on the toss of a die is 2/6 or 1/3.

In many cases, however, we are not able to enumerate all the possible ways in which an event can occur, and, therefore, we use the *relative frequency definition of probability*. This is defined as the number of times that the event of interest has occurred divided by the total number of trials (or opportunities for the event to occur). Since it is based on previous data, it is called the *a posteriori definition of probability*.

For instance, if you select at random a white American female, the probability of her dying of heart disease is 0.00199. This is based on the finding that per 100,000 white American females, 199 died of coronary heart disease (estimates are for 2011, National Vital Statistics System, National Center for Health Statistics, Centers for Disease Control and Prevention). When you consider the probability of a white American female who is between ages 45 and 54, the figure drops to 0.00041 (or 41 women in that age group out of 100,000), and when you consider women 75–84

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_2

years old, the figure rises to 0.00913 (or 913 per 100,000). For white men 75–84 years old, it is 0.01434 (or 1,434 per 100,000). The two important points are (1) to determine a probability, *you must specify the population to which you refer*, for example, all white females, white males between 65 and 74, nonwhite females between 65 and 74, and so on, and (2) the *probability figures are constantly revised* as new data become available.

This brings us to the notion of *expected frequency*. If the probability of an event is P and there are N trials (or opportunities for the event to occur), then we can expect that the event *will* occur $N \times P$ times. It is necessary to remember that probability "works" for large numbers. When in tossing a coin we say the probability of it landing on heads is 0.50, we mean that in many tosses half the time the coin will land heads. If we toss the coin ten times, we may get three heads (30%) or six heads (60%), which are a considerable departure from the 50% we expect. But if we toss the coin 200,000 times, we are very likely to be close to getting exactly 100,000 heads or 50%.

Expected frequency is really the way in which probability "works." It is difficult to conceptualize applying probability to an individual. For example, when TV announcers proclaim there will be, say, 400 fatal accidents in State X on the fourth of July, it is impossible to say whether any individual person will in fact have such an accident, but we can be pretty certain that the number of such accidents will be very close to the predicted 400 (based on probabilities derived from previous fourth of July statistics).

2.2 Combining Probabilities

There are two laws for combining probabilities that are important. First, if there are *mutually exclusive events* (i.e., if one occurs, the other cannot), the probability of either one or the other occurring is the *sum* of their individual probabilities. Symbolically,

$$P(A \ \ or \ \ B) = P(A) + P(B)$$

An example of this is as follows: The probability of getting either a 3 or a 4 on the toss of a die is 1/6 + 1/6 = 2/6.

A useful thing to know is that the sum of the individual probabilities of all possible mutually exclusive events must equal 1. For example, if A is the event of winning a lottery, and not A (written as $\overline{A}$) is the event of not winning the lottery, then $\left(P(A) + \left(\overline{A}\right) = 1.0\right)$ and $P\left(\overline{A}\right) + (A) = 1 - P(A)$.

Second, if there are two independent events (i.e., the occurrence of one is not related to the occurrence of the other), the joint probability of their occurring together (jointly) is the *product* of the individual probabilities. Symbolically,

$$P(A \ \ and \ \ B) = P(A) \times P(B)$$

An example of this is the probability that on the toss of a die you will get a number that is both even and divisible by 3. This probability is equal to 1/2 × 1/3 = 1/6. (The only number both even and divisible by 3 is the number 6.)

The joint probability law is used to test whether events are independent. If they are independent, the product of their individual probabilities should equal the joint probability. If it does not, they are not independent. It is the basis of the chi-square test of significance, which we will consider in the next section.

Let us apply these concepts to a medical example. The mortality rate for those with a heart attack in a special coronary care unit in a certain hospital is 15 %. Thus, the probability that a patient with a heart attack admitted to this coronary care unit will die is 0.15 and that he will survive is 0.85. If two men are admitted to the coronary care unit on a particular day, let *A* be the event that the first man dies and let *B* be the event that the second man dies.

The probability that both will die is

$$P(A \ \ and \ \ B) = P(A) \times (B) = .15 \times .15 = .0225$$

We assume these events are independent of each other, so we can multiply their probabilities. Note, however, that the probability that either one *or* the other will die from the heart attack is *not* the sum of their probabilities because these two events are not mutually exclusive. It is possible that both will die (i.e., both *A* and *B* can occur).

To make this clearer, a good way to approach probability is through the use of Venn diagrams, as shown in Figure 2.1. Venn diagrams consist of squares that represent the universe of possibilities and circles that define the events of interest.

In diagrams 1, 2, and 3, the space inside the square represents all *N* possible outcomes. The circle marked A represents all the outcomes that constitute event *A*; the circle marked B represents all the outcomes that constitute event *B*. Diagram 1 illustrates two mutually exclusive events; an outcome in circle A cannot also be in circle B. Diagram 2 illustrates two events that can occur jointly: an outcome in circle A can also be an outcome belonging to circle B. The shaded area marked AB represents outcomes that are the occurrence of both *A and B*. The diagram 3 represents two events where one (*B*) is a subset of the other (*A*); an outcome in circle B must also be an outcome constituting event *A*, but the reverse is not necessarily true.

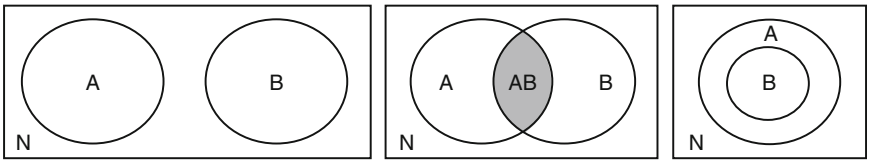


Figure 2.1 Venn diagrams

It can be seen from diagram 2 that if we want the probability of an outcome being either *A or B* and if we add the outcomes in circle A to the outcomes in circle B, we have added in the outcomes in the shaded area twice. Therefore, we must subtract the outcomes in the shaded area (A *and B*) also written as (*AB*) once to arrive at the correct answer. Thus, we get the result

$$P(A \ \ or \ \ B) = P(A) + P(B) - P(AB)$$

2.3 Conditional Probability

Now let us consider the case where the chance that a particular event happens is dependent on the outcome of another event. The probability of *A*, given that *B* has occurred, is called the conditional probability of *A* given *B* and is written symbolically as *P*(*A*|*B*). An illustration of this is provided by Venn diagram 2. When we speak of conditional probability, the denominator becomes all the outcomes in circle B (instead of all *N* possible outcomes) and the numerator consists of those outcomes that are in that part of *A*, which also contains outcomes belonging to *B*. This is the shaded area in the diagram labeled AB. If we return to our original definition of probability, we see that

$$P(A|B) = \frac{n_{AB}}{n_B}$$

(the number of outcomes in both *A and B*, divided by the total number of outcomes in *B*).

If we divide both numerator and denominator by *N*, the total number of *all* possible outcomes, we obtain

$$P(A|B) = \frac{n_{AB}/N}{n_B/N} = \frac{P(A \ \ and \ \ B)}{P(B)}$$

Multiplying both sides by *P*(*B*) gives the *complete* multiplicative law:

$$P(A \ \ and \ \ B) = P(A|B) \times P(B)$$

Of course, if *A* and *B* are independent, then the probability of *A* given *B* is just equal to the probability of *A* (since the occurrence of *B* does not influence the occurrence of *A*) and we then see that

$$P(A \ \ and \ \ B) = P(A) \times P(B)$$

2.4 Bayesian Probability

Imagine that M is the event "loss of memory" and B is the event "brain tumor." We can establish from research on brain tumor patients the probability of *memory loss given a brain tumor,* $P(M|B)$. A clinician, however, is more interested in the probability of *a brain tumor, given that a patient has memory loss,* $P(B|M)$.

It is difficult to obtain that probability directly because one would have to study the vast number of persons with memory loss (which in most cases comes from other causes) and determine what proportion of them have brain tumors.

Bayes' equation (or Bayes' theorem) estimates $P(B|M)$. as follows:

$$P(brain\ \ tumor,\ \ given\ \ memory\ \ loss) = \frac{P(memory\ \ loss,\ \ given\ \ brain\ \ tumor) \times P(brain\ \ tumor)}{P(memory\ \ loss)}$$

In the denominator, the event of "memory loss" can occur either among people with brain tumor, with probability $= P(M|B)P(B)$, or among people with no brain tumor, with probability $= \mathrm{P}\left(M|\overline{B}\right)P\left(\overline{B}\right)$.

Thus,

$$P(B|M) = \frac{P(M|B)P(B)}{P(M|B)P(B) + P\left(M|\overline{B}\right)P\left(\overline{B}\right)}$$

The overall probability of a brain tumor, $P(B)$, is the "a priori probability," which is a sort of "best guess" of the prevalence of brain tumors.

2.5 Odds and Probability

When the odds of a particular horse *losing* a race are said to be 4–1, he has a $4/5 = 0.80$ probability of losing. To convert an odds statement to probability, we add $4 + 1$ to get our denominator of 5. The odds of the horse *winning* are 1–4, which means he has a probability of winning of $1/5 = 0.20$:

$$The\ \ odds\ \ in\ \ favour\ \ of\ \ A = \frac{P(A)}{P(not\ \ A)} = \frac{P(A)}{1 - P(A)}$$

$$P(A) = \frac{odds}{1 + odds}$$

The odds of drawing an ace $= 4$ (aces in a deck) to 48 (cards that are not aces) $= 1$–12; therefore, $P(\text{ace}) = 1/13$. The odds *against* drawing an ace $= 12$–1; $P(\text{not ace}) = 12/13$.

In medicine, odds are often used to calculate an *odds ratio*. An odds ratio is simply the ratio of two odds. For example, say that in a particular study comparing lung cancer patients with controls, it was found that the odds of being a lung cancer case for people who smoke were 5–4 (5/4) and the odds of having lung cancer for nonsmokers was 1–8 (1/8), then the odds ratio would be

$$\frac{5/4}{1/8} = \frac{5 \times 8}{4 \times 1} = \frac{40}{4} = 10$$

An odds ratio of 10 means that the odds of being a lung cancer case is ten times greater for smokers than for nonsmokers.

Note, however, that we cannot determine from such an analysis what the probability of getting lung cancer is for smokers, because in order to do that we would have to know how many people out of all smokers developed lung cancer, and we haven't studied all smokers; all we do know is how many out of all our lung cancer cases were smokers. Nor can we get the probability of lung cancer among nonsmokers, because we would have to a look at a population of nonsmokers and see how many of them developed lung cancer. All we do know is that smokers have tenfold greater odds of having lung cancer than nonsmokers.

More on this topic is presented in Section 4.13.

2.6 Likelihood Ratio

A related concept is the likelihood ratio (LR), which tells us how likely it is that a certain result would arise from one set of circumstances in relation to how likely the result would arise from an opposite set of circumstances.

For example, if a patient has a sudden loss of memory, we might want to know the likelihood ratio of that symptom for a brain tumor, say. What we want to know is the likelihood that the memory loss arose out of the brain tumor *in relation to* the likelihood that it arose from some other condition. The likelihood ratio is a ratio of conditional probabilities.

$$\begin{aligned} LR &= \frac{P(\textit{memory loss, given brain tumor})}{P(\textit{memory loss, given no brain tumor})} \\ &= \frac{P(M \textit{ given } B)}{P(M \textit{ given not } B)} \end{aligned}$$

Of course to calculate this LR, we would need to have estimates of the probabilities involved in the equation; that is, we would need to know the following: among persons who have brain tumors, what proportion have memory loss, and among persons who don't have brain tumors, what proportion have memory loss. It may sometimes be quite difficult to establish the denominator of the likelihood ratio

because we would need to know the prevalence of memory loss in the general population.

The LR is perhaps more practical to use than the Bayes theorem, which gives the probability of a particular disease given a particular symptom. In any case, it is widely used in variety of situations because it addresses this important question: If a patient presents with a symptom, what is the likelihood that the symptom is due to a particular disease *rather than* to some other reason than this disease?

2.7 Summary of Probability

- ***Additive Law***

$$P(A \ \ or \ \ B) = P(A) + P(B) - P(A \ \ and \ \ B)$$

If events are mutually exclusive: $P\ (A\ or\ B) = P\ (A) + P\ (B)$.

- ***Multiplicative Law***

$$P(A \ \ and \ \ B) = P(A|B) \times P(B)$$

If events are independent: $P(A\ and\ B) = P(A) \times P(B)$.

- ***Conditional Probability***

$$P(A|B) = \frac{P(A \ \ and \ \ B)}{P(B)}$$

- ***Bayes' Theorem***

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P\left(A|\overline{B}\right)P\left(\overline{B}\right)}$$

- ***Odds of A***

$$\frac{P(A)}{1 - P(A)}$$

- ***Likelihood Ratio***

$$\frac{P(A|B)}{P\left(A|\overline{B}\right)}$$

Chapter 3
Mostly About Statistics

3.1 Chi-Square for 2 × 2 Tables

The chi-square test is a statistical method to determine whether the results of an experiment may arise by chance or not. Let us, therefore, consider the example of testing an anticoagulant drug on female patients with myocardial infarction. We hope the drug lowers mortality, but we set up our null hypothesis as follows:

- Null hypothesis: There is no difference in mortality between the treated group of patients and the control group
- Alternate hypothesis: The mortality in the treated group is lower than in the control group

(The data for our example come from a study done a long time ago and refer to a specific high-risk group.[8] They are used for illustrative purposes, and they do not reflect current mortality rates for people with myocardial infarction.)

We then record our data in a 2 × 2 *contingency* table in which each patient is classified as belonging to one of the four cells:

Observed frequencies

	Control	Treated	
Lived	89	223	312
Died	40	39	79
Total	129	262	391

The mortality in the control group is 40/129 = 31% and in the treated it is 39/262 = 15%. But could this difference have arisen by chance? We use the chi-square test to answer this question. What we are really asking is whether the two categories of classification (control vs. treated by lived vs. died) are independent of

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_3

each other. If they are independent, what frequencies would we expect in each of the cells? And *how different are our observed frequencies from the expected ones?* How do we measure the size of the difference?

To determine the expected frequencies, consider the following:

	Control	Treated	
Lived	a	b	$(a + b)$
Died	c	d	$(c + d)$
Total	$(a + c)$	$(b + d)$	N

If the categories are independent, then the probability of a patient being both a control and living is $P(\text{control}) \times P(\text{lived})$. [Here we apply the law referred to in Chapter 2 on the joint probability of two independent events.]

The expected frequency of an event is equal to the probability of the event times the number of trials $= N \times P$. So the expected number of patients who are *both* controls *and* live is

$$N \times P(control \ \ and \ \ lived) = N \times P(control) \times P(lived)$$
$$= N\left[\frac{(a+c)}{N} \times \frac{(a+b)}{N}\right] = (a+c) \times \frac{(a+b)}{N}$$

In our case, this yields the following table:

	Control	Treated	
Lived	$129 \times \frac{312}{391} = 103$	$262 \times \frac{312}{391} = 209$	312
Died	$129 \times \frac{79}{391} = 26$	$262 \times \frac{79}{391} = 53$	79
Total	129	262	391

Another way of looking at this is to say that since 80% of the patients in the total study lived (i.e., 312/391 = 80%), we would expect that 80% of the control patients and 80% of the treated patients would live. These expectations differ, as we see, from the observed frequencies noted earlier; that is, those patients treated did, in fact, have a lower mortality than those in the control group.

Well, now that we have a table of observed frequencies and a table of expected values, how do we know just how different they are? Do they differ just by chance or is there some other factor that causes them to differ? To determine this, we calculate a value called chi-square (also written as χ^2). This is obtained by taking the observed value in each cell, subtracting from it the expected value in each cell, squaring this difference, and dividing by the expected value for each cell. When this is done for

each cell, the four resulting quantities are added together to give a number called chi-square. Symbolically this formula is as follows:

$$\frac{(O_a - e_a)^2}{e_a} + \frac{(O_b - e_b)^2}{e_b} + \frac{(O_c - e_c)^2}{e_c} + \frac{(O_d - e_d)^2}{e_d}$$

where O is the observed frequency and e is the expected frequency in each cell.

This number, called chi-square, is a statistic that has a known distribution. What that means, in essence, is that for an infinite number of such 2 × 2 tables, chi-squares have been calculated, and we thus know what the probability is of getting certain values of chi-square. Thus, when we calculate a chi-square for a particular 2 × 2 contingency table, we know how likely it is that we could have obtained a value as large as the one that we actually obtained strictly by chance, under the assumption the hypothesis of independence is the correct one, that is, if the two categories of classification were unrelated to one another or if the null hypothesis were true. The particular value of chi-square that we get for our example happens to be 13.94.

From our knowledge of the distribution of values of chi-square, we know that if our null hypothesis is true, that is, if there is no difference in mortality between the control and treated group, then the probability that we get a value of chi-square as large or larger than 13.94 by chance alone is very, very low; in fact this probability is less than 0.005. Since it is not likely that we would get such a large value of chi-square by chance under the assumption of our null hypothesis, *it must be that it has arisen not by chance but because our null hypothesis is incorrect.* We, therefore, reject the null hypothesis at the 0.005 level of significance and accept the alternate hypothesis; that is, we conclude that among women with myocardial infarction, the new drug does reduce mortality. The probability of obtaining these results by chance alone is less than 5/1000 (.005). Therefore, the probability of rejecting the null hypothesis, when it is in fact true (type I error), is less than 0.005.

The probabilities for obtaining various values of chi-square are tabled in most standard statistics texts, so that the procedure is to calculate the value of chi-square and then look it up in the table to determine whether or not it is significant. That value of chi-square that must be obtained from the data in order to be significant is called the *critical value*. The critical value of chi-square at the 0.05 level of significance for a 2 × 2 table is 3.84. This means that when we get a value of 3.84 *or greater* from a 2 × 2 table, we can say there is a significant difference between the two groups. Appendix 1 provides some critical values for chi-square and for other tests.

In actual usage, a correction is applied for 2 × 2 tables known as the Yates' correction and calculation is done using the formula

$$\frac{N\left[|ad - bc| - \frac{N}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

Note: $|ad–bc|$ means the absolute value of the difference between $a \times d$ and $b \times c$. In other words, if $a \times d$ is greater than $b \times c$, subtract bc from ad; if bc is greater than ad, subtract ad from bc. The corrected chi-square so calculated is 12.95, still well above the 3.84 required for significance.

The chi-square test should not be used if the numbers in the cells are too small. The rules of thumb: When the total N is greater than 40, use the chi-square test with Yates' correction. When N is between 20 and 40 and the expected frequency in each of the four cells is 5 or more, use the corrected chi-square test. If the smallest expected frequency is less than 5, or if N is less than 20, use Fisher's test.

While the chi-square test approximates the probability, Fisher's exact test gives the exact probability of getting a table with values like those obtained or even more extreme. A sample calculation is shown in Appendix 2. The calculations are unwieldy, but Fisher's exact test is also usually included in most statistics programs for personal computers. More on this topic may be found in the book *Statistical Methods for Rates and Proportions* by Joseph L. Fleiss. The important thing is to know when the chi-square test is or is not appropriate.

3.2 McNemar Test

Suppose we have the situation where measurements are made on the same group of people before and after some intervention, or suppose we are interested in the agreement between two judges who evaluate the same group of patients on some characteristics. In such situations, the before and after measures, or the opinions of two judges, are not independent of each other, since they pertain to the same individuals. Therefore, the chi-square test or Fisher's exact test is not appropriate. Instead, we can use the McNemar test.

Consider the following example. Case histories of patients who were suspected of having ischemic heart disease (a decreased blood flow to the heart because of clogging of the arteries) were presented to two cardiology experts. The doctors were asked to render an opinion on the basis of the available information about the patient. They could recommend either (1) that the patient should be on medical therapy or (2) that the patient have an angiogram, which is an invasive test, to determine if the patient is a suitable candidate for coronary artery bypass graft surgery (known as CABG). Table 3.1 shows the results of these judgments on 661 patients.

Note that in cell b Expert 1 advised surgery and Expert 2 advised medical therapy for 97 patients, whereas in cell c Expert 1 advised medical therapy and Expert 2 advised surgery for 91 of the patients. Thus, the two physicians disagree in 188 of the 661 cases or 28% of the time. Cells a and d represent patients about whom the two doctors agree. They agree in 473 out the 661 cases or 72% of the time.

To determine whether the observed disagreement could have arisen by chance alone under the null hypothesis of no real disagreement in recommendations between the two experts, we calculate a type of chi-square value as follows:

Table 3.1 Calculating Chi-square

		EXPERT 1		
		Medical	Surgical	
EXPERT 2	Medical	$a = 397$	$b = 97$	$a + b = 494$
	Surgical	$c = 91$	$d = 76$	$c + d = 167$
		$a + c = 488$	$b + d = 173$	$N = 661$

$$\chi^2(chi-square) = \frac{(|b-c|-1|)^2}{(b+c)} = \frac{25}{188} = .13$$

(|*b*–*c*| means the absolute value of the difference between the two cells, that is, irrespective of the sign; the −1 in the numerator is analogous to the Yates' correction for chi-square described in Section 3.1 and gives a better approximation to the chi-square distribution.) A chi-square of 0.13 does not reach the critical value of chi-square of 3.84 needed for a 0.05 significance level, as described in Section 3.1, so we cannot reject the null hypothesis, and we conclude that our data are consistent with no difference in the opinions of the two experts. Were the chi-square test significant, we would have to reject the null hypothesis and say the experts significantly disagree. However, such a test does not tell us about the *strength* of their agreement, which can be evaluated by a statistic called Kappa.

3.3 Kappa

The two experts could be agreeing just by chance alone, since both experts are more likely to recommend medical therapy for these patients. Kappa is a statistic that tells us the extent of the agreement between the two experts above and beyond chance agreement.

$$K = \frac{Proportion\ of\ observed\ agreement - Proportion\ of\ agreement\ by\ chance}{1 - Proportion\ of\ agreement\ by\ chance}$$

To calculate the expected number of cases in each cell of the table, we follow the procedure described for chi-square in Section 3.1. The cells a and d in Table 3.1 represent agreement. The expected number by chance alone is

$$cell\ \ a: \frac{494 \times 488}{661} = 365$$
$$cell\ \ d: \frac{167 \times 173}{661} = 44$$

So the proportion of agreement expected by chance alone is

$$\frac{365 + 44}{661} = .619$$

that is, by chance alone, the experts would be expected to agree 62% of the time. The proportion of observed agreement is

$$\frac{397 + 76}{661} = .716$$
$$Kappa = \frac{.716 - .619}{1 - .619} = \frac{.097}{.381} = .25$$

If the two experts agreed at the level of chance only, Kappa would be 0; if the two experts agreed perfectly, Kappa would be 1.

3.4 Description of a Population: Use of the Standard Deviation

In the case of continuous variables, as for discrete variables, we may be interested in description or in inference. When we wish to describe a population with regard to some characteristic, we generally use the mean or average as an index of *central tendency* of the data.

Other measures of central tendency are the *median* and the *mode*. The median is that value above which 50% of the other values lie and below which 50% of the values lie. It is the middle value or the 50th percentile. To find the median of a set of scores, we arrange them in ascending (or descending) order and locate the middle value if there are an odd number of scores or the average between the two middle scores if there are an even number of scores. The mode is the value that occurs with the greatest frequency. There may be several modes in a set of scores but only one median and one mean value. These definitions are illustrated below. The mean is the measure of central tendency most often used in inferential statistics.

Measures of central tendency	
Set of scores	Ordered
12	6
12	8
6	10
8	**11 Median**
11	*12 Mode*
10	***12***
15	15
SUM: 74	*Mean* = 74/7 = 10.6

The true mean of the population is called *m*, and we estimate that mean from data obtained from a sample of the population. The sample mean is called $\overline{x}$ (read as x bar). We must be careful to specify exactly the population from which we take a sample. For instance, in the general population, the average IQ is 100, but the average IQ of the population of children aged 6–11 years whose fathers are college graduates is 112.[9] Therefore, if we take a sample from either of these populations, we would be estimating a different population mean, and we must specify to which population we are making inferences.

However, the mean does not provide an adequate description of a population. What is also needed is some measure of *variability* of the data around the mean. Two groups can have the same mean but be very different. For instance, consider a hypothetical group of children each of whose individual IQ is 100; thus, the mean is 100. Compare this to another group whose mean is also 100 but includes individuals with IQs of 60 and those with IQs of 140. Different statements must be made about these two groups: One is composed of all average individuals and the other includes both retardates and geniuses.

The most commonly used index of variability is the *standard deviation (s.d.)*, which is a type of measure related to the average distance of the scores from their mean value. The square of the standard deviation is called *variance*. The population standard deviation is denoted by the Greek letter σ (sigma). When it is calculated from a *sample*, it is written as s.d. and is illustrated in the example below:

IQ scores		Deviations from mean		Squared scores for B
Group A	Group B	$x_i - \overline{x}_B$	$(x_i - \overline{x}_B)^2$	x_b^2
100	60	−40	1,600	3,600
100	140	40	1,600	19,600
100	80	−20	400	6,400
100	120	20	400	14,400
$\Sigma = 400 \overline{x}_A = \text{mean} = 100$	$\Sigma = 400\ \overline{x}_B = \text{mean} = 100$	$\Sigma = 0$	$\Sigma = 4{,}000$ of squared deviations	$\Sigma = 44{,}000$ sum of squares

Note: The symbol "Σ" means "sum."
Note: The sum of deviations from the mean, as in column 3, is always 0; that is why we sum the squared deviations, as in column 4.

$$\bar{x}_A = mean = \frac{400}{4} = 100; \quad \bar{x}_B = \frac{400}{4} = 100$$

$$s.d. = \sqrt{\frac{\Sigma \ of(each \ \ value - mean \ \ of \ \ group)^2}{n-1}} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

$$s.d._A = \frac{0}{3} = 0;$$

(In group A since each score is equal to the mean of 100, there are no deviations from the mean of A.)

$$s.d._B = \sqrt{\frac{4,000}{3}} = \sqrt{1,333} = 36.51$$

An equivalent formula for s.d. that is more suited for actual calculations is

$$s.d. = \sqrt{\frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}}$$

For group B we have

$$s.d. = \sqrt{\frac{44,000 - 4(100)^2}{3}} = \sqrt{\frac{44,000 - 40,000}{3}} = \sqrt{\frac{4,000}{3}} = 36.51$$

$Variance = (s.d.)^2$

Note the mean of both groups is 100, but the standard deviation of group A is 0, while the s.d. of group B is 36.51. (We divide the squared deviations by n–1 rather than by n because we are estimating the population σ from sample data, and dividing by n–1 gives a better estimate. The mathematical reason is complex and beyond the scope of this book.)

3.5 Meaning of the Standard Deviation: The Normal Distribution

The standard deviation is a measure of the dispersion or spread of the data. Consider a variable like IQ, which is normally distributed; that is, it can be described by the familiar, bell-shaped curve where most of the values fall around the mean with decreasing number of values at either extremes. In such a case, 68% of the values lie within 1 standard deviation on either side of the mean, 95% of the values lie within 2 standard deviations of the mean, and 99% of the values lie within 3 standard deviations of the mean. (The IQ test was originally constructed so that it had a mean of 100 and a standard deviation of 16.)

This is illustrated in Figure 3.1.

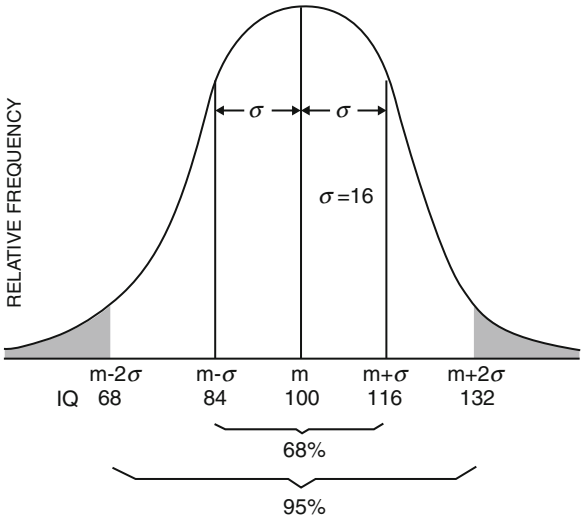


Figure 3.1 Normal distribution

In the population at large, 95% of people have IQs between 68 and 132. Approximately 2.5% of people have IQs above 132 and another 2.5% of people have IQs below 68. (This is indicated by the shaded areas at the tails of the curves.)

If we are estimating from a sample and if there are a large number of observations, the standard deviation can be estimated from the *range of the data*, that is, the difference between the smallest and the highest value. Dividing the range by 6 provides a rough estimate of the standard deviation if the distribution is normal, because 6 standard deviations (3 on either side of the mean) encompass 99%, or virtually all, of the data.

On an individual, clinical level, knowledge of the standard deviation is very useful in deciding whether a laboratory finding is normal, in the sense of "healthy." Generally a value that is more than 2 standard deviations away from the mean is suspect, and perhaps further tests need to be carried out.

For instance, suppose as a physician you are faced with an adult male who has a hematocrit reading of 39. Hematocrit is a measure of the amount of packed red cells in a measured amount of blood. A low hematocrit may imply anemia, which in turn may imply a more serious condition. You also know that the average hematocrit reading for adult males is 47. Do you know whether the patient with a reading of 39 is normal (in the sense of health) or abnormal? You need to know the standard deviation of the distribution of hematocrits in people before you can determine whether 39 is a normal value. In point of fact, the standard deviation is approximately 3.5; thus, plus or minus 2 standard deviations around the mean results in the range of from 40 to 54 so that 39 would be slightly low. For adult females, the mean hematocrit is 42 with a standard deviation of 2.5, so that the range of plus or minus 2 standard deviations away from the mean is from 37 to 47. Thus, if an adult female came to you with a hematocrit reading of 39, she would be considered in the "normal" range.

3.6 The Difference Between Standard Deviation and Standard Error

Often data in the literature are reported as ± s.d. (read as mean + or −1 standard deviation). Other times they are reported as ± s.e. (read as mean + or −1 standard error). *Standard error* and *standard deviation* are often confused, but they serve quite different functions. To understand the concept of standard error, you must remember that the purpose of statistics is to draw inferences from samples of data to the population from which these samples came. Specifically, we are interested in estimating the true mean of a population for which we have a sample mean based on, say, 25 cases. Imagine the following:

Population IQ scores, x_i	Sample means based on 25 people randomly selected
110	$\overline{x_1} = 102$
100	
105	$\overline{x_2} = 99$
98	
140	$\overline{x_3} = 101$
–	$\overline{x_4} = 98$
–	–
100	100
m = mean of all the x_is	$m_{\bar{x}} = m$, mean of the means is m, the population mean
σ = population standard deviation	$\frac{\sigma}{\sqrt{n}}$ = Standard deviation of the distribution of the $\bar{x}$ called the standard error of the mean = $\sigma_{\bar{x}}$

There is a population of IQ scores whose mean is 100 and its standard deviation is 16. Now imagine that we draw a sample of 25 people at random from that population and calculate the sample mean $\bar{x}$. This sample mean happens to be 102. If we took another sample of 25 individuals, we would probably get a slightly different sample mean, for example, 99. Suppose we did this repeatedly an infinite (or a very large) number of times, each time throwing the sample, we just drew back into the population pool from which we would sample 25 people again. We would then have a very large number of such sample means. These sample means would form a normal distribution. Some of them would be very close to the true population mean of 100, and some would be at either end of this "distribution of means" as in Figure 3.2.

This distribution of sample means would have its own standard deviation, that is, a measure of the spread of the data around the mean of the data. In this case, the data are sample means rather than individual values. The standard deviation of this distribution of means is called the *standard error of the mean*.

It should be pointed out that this distribution of means, which is also called the sampling distribution of means, is a theoretical construct. Obviously, we don't go around measuring samples of the population to construct such a distribution. Usually, in fact, we just take *one sample of 25* people and imagine what this

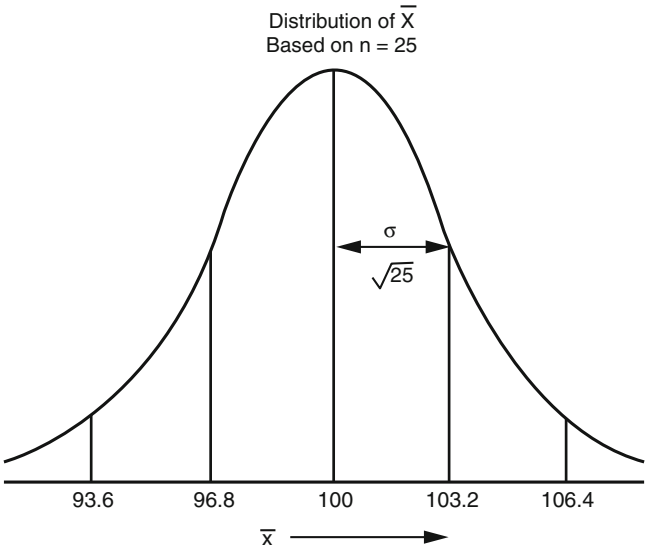


Figure 3.2 Distribution of sample means

distribution might be. However, due to certain mathematical derivations, we know a lot about this theoretical distribution of population means, and therefore we can draw important inferences based on just one sample mean. What we do know is that the distribution of means is a normal distribution, that its mean is the same as the population mean of the individual values (i.e., *the mean of the means is m*), and that its standard deviation is equal to the standard deviation of the original individual values divided by the square root of the number of people in the sample.

Standard error of the mean =

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

In this case it would be

$$\frac{16}{\sqrt{25}} = \frac{16}{5} = 3.2$$

The distribution of means would look as shown in Figure 3.2.

Please note that when we talk about population values, which we usually don't know but are trying to estimate, we refer to the mean as *m* and the standard deviation as σ. When we talk about values calculated from samples, we refer to the mean as $\overline{x}$, the standard deviation as s.d., and the standard error as s.e.

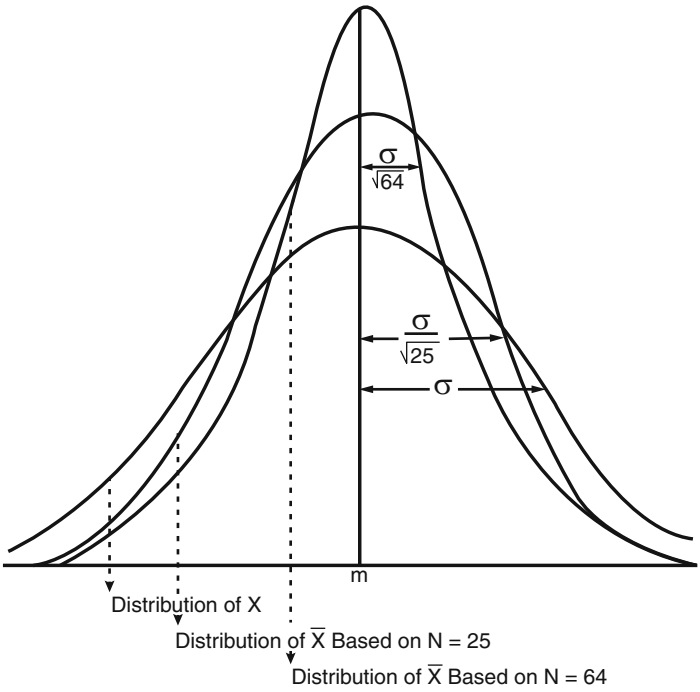


Figure 3.3 Distribution of means for different sample sizes

Now imagine that we have a distribution of means based on samples of 64 individuals. The mean of these means is also *m*, but its dispersion, or standard error, is smaller. It is $16/\sqrt{64} = 16/8 = 2$. This is illustrated in Figure 3.3.

It is easily seen that if we take a sample of 25 individuals, their mean is likely to be closer to the true mean than the value of a single individual, and if we draw a sample of 64 individuals, their mean is likely to be even closer to the true mean than was the mean we obtained from the sample of 25. Thus, the larger the sample size, the better is our estimate of the true population mean.

The standard deviation is used to describe the dispersion or variability of the scores. The standard error is used to draw inferences about the population mean from which we have a sample. We draw such inferences by constructing confidence intervals, which are discussed in Section 3.11.

3.7 Standard Error of the Difference Between Two Means

This concept is analogous to the concept of standard error of the mean. The standard error of the differences between two means is the standard deviation of a theoretical distribution of differences between two means. Imagine a group of men and a group of women each of whom have an IQ measurement. Suppose we take a sample of 64 men and a sample of 64 women, calculate the mean IQs of these two samples, and obtain their differences. If we were to do this an infinite number of times, we would get a *distribution of differences* between sample means of two groups of 64 each. These difference scores would be normally distributed; their mean would be the true average difference between the populations of men and women (which we are trying to infer from the samples), and the standard deviation of this distribution is called the *standard error of the differences between two means*.

The standard error of the difference between two means of populations X and Y is given by the formula

$$\sigma_{\overline{x}-\overline{y}} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

where ${\sigma_x}^2$ is the variance of population X and ${\sigma_y}^2$ is the variance of population Y, n_x is the number of cases in the sample from population X, and n_y is the number of cases in the sample from population Y.

In some cases, we know or assume that the variances of the two populations are equal to each other and that the variances that we calculate from the samples we have drawn are both estimates of a common variance. In such a situation, we would want to pool these estimates to get a better estimate of the common variance. We denote this *pooled estimate* as ${s_{\text{pooled}}}^2 = {s_p}^2$, and we calculate the standard error of the difference between means as

$$s.e._{\overline{x}-\overline{y}} = \sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)} = s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

We calculate ${s_p}^2$ from sample data:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

This is the equivalent to

$$s_p^2 = \frac{\Sigma(x_i - \overline{x})^2 + \Sigma(y_i - \overline{y})^2}{n_x + n_y - 2}$$

Since in practice we will always be calculating our values from sample data, we will henceforth use the symbology appropriate to that.

3.8 Z Scores and the Standardized Normal Distribution

The standardized normal distribution is one whose mean = 0, standard deviation = 1, and the total area under the curve = 1. The standard normal distribution looks like the one shown in Figure 3.4.

On the abscissa, instead of x, we have a transformation of x called the standard score; Z. Z is derived from x by the following:

$$Z = \frac{x - m}{\sigma}$$

Thus, the Z score really tells you how many standard deviations from the mean a particular x score is.

Any distribution of a normal variable can be transformed to a distribution of Z by taking each x value, subtracting from it the mean of x (i.e., m), and dividing this deviation of x from its mean, by the standard deviation. Let us look at the IQ distribution again in Figure 3.5.

Thus, an IQ score of 131 is equivalent to a Z score of 1.96 (i.e., it is 1.96, or nearly 2, standard deviations above the mean IQ).

$$Z = \frac{131 - 100}{16} = 1.96$$

One of the nice things about the Z distribution is that the probability of a value being anywhere between two points is equal to the area under the curve between those two points. (Accept this on faith.) It happens that the area to the left of 1.96 corresponds to a probability of 0.025, or 2.5% of the total curve. Since the curve is symmetrical, the probability of Z being to the left of −1.96 is also 0.025. Invoking the additive law of probability (Section 2.2), the probability of a Z being *either* to the left of −1.96 *or* to the left of +1.96 is 0.025 + 0.025 = .05. Transforming back up to

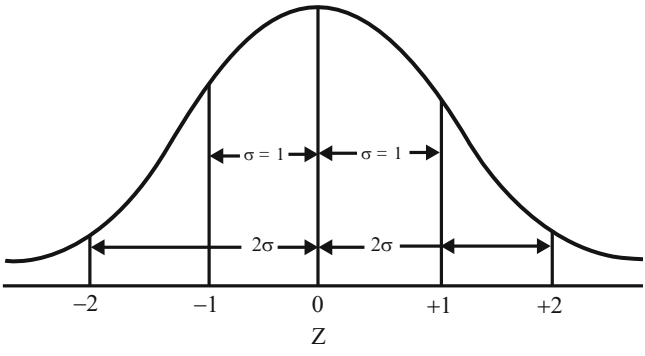


Figure 3.4 Standard normal distribution

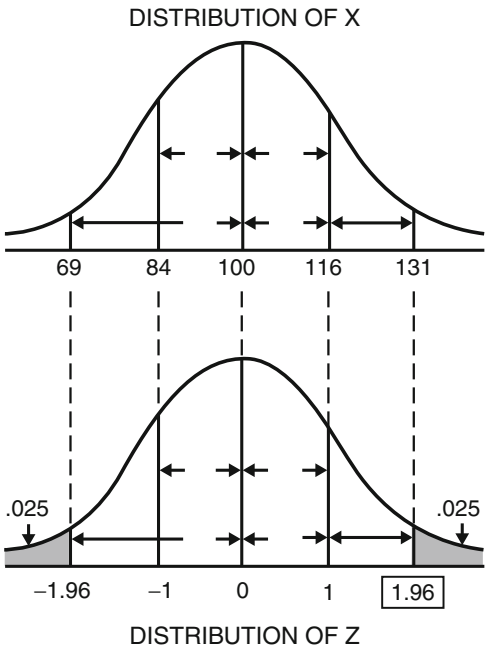


Figure 3.5 Distribution of Z scores

x, we can say that the probability of someone having an IQ outside of 1.96 standard deviations away from the mean (i.e., above 131 or below 69) is 0.05, or only 5% of the population have values that extreme. (Commonly, the Z value of 1.96 is rounded off to ±2 standard deviations from the mean as corresponding to the cutoff points beyond which lies 5% of the curve, but the accurate value is 1.96.)

A very important use of Z derives from the fact that we can also convert a sample mean (rather than just a single individual value) to a Z score.

$$Z = \frac{\overline{x} - m}{\sigma_{\overline{x}}}$$

The numerator now is the distance of the sample mean from the population mean, and the denominator is the standard deviation of the distribution of means, which is the *standard error of the mean*. This is illustrated in Figure 3.6, where we are considering means based on 25 cases each. The s.e. is $16/\sqrt{25} = 16/5 = 3.2$.

Now we can see that a sample mean of 106.3 corresponds to a Z score of 1.96.

$$Z = \frac{106.3 - 100}{3.2} = 1.96$$

We can now say that the probability that the *mean IQ of a group of 25 people* is greater than 106.3 is 0.025. The probability that such a mean is less than 93.7 is also 0.025.

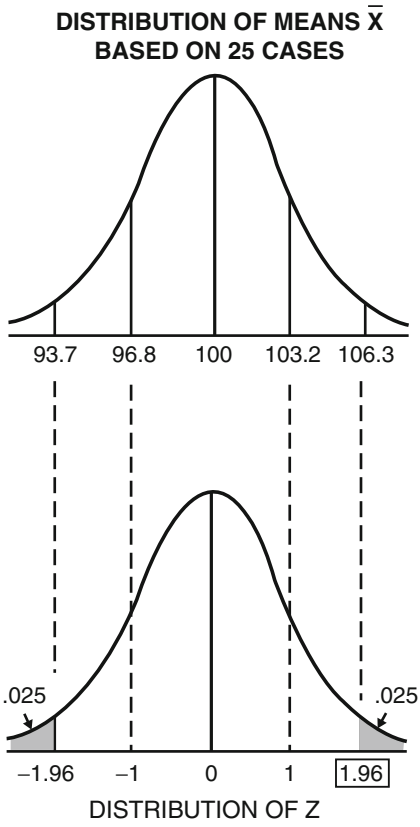


Figure 3.6 Distribution of means based on samples of size 25

A Z score can also be calculated for the *difference between two means*.

$$Z = \frac{(\bar{x}_A - \bar{x}_B) - (m_A - m_B)}{\sigma_{\bar{x}_A - \bar{x}_B}}$$

But m_A–m_B is commonly hypothesized to be 0 so the formula becomes

$$Z = \frac{\bar{x}_A - \bar{x}_B}{\sigma_{\bar{x}_A - \bar{x}_B}}$$

You can see that a Z score in general is a distance between some value and its mean divided by an appropriate standard error.

This becomes very useful later on when we talk about confidence intervals in Sections 3.10, 3.11, 3.12, 3.13, 3.14 and 3.15.

3.9 The t Statistic

Suppose we are interested in sample means and we want to calculate a Z score. We don't know what the population standard deviation is, but if our samples are very large, we can get a good estimate of σ by calculating the standard deviation, s.d., from our sample, and then getting the standard error as usual: s.e. = s.d./$\sqrt{n}$. But often our sample is not large enough. We can still get a standardized score by calculating a value called Student's t:

$$t = \frac{\overline{x} - m}{s.e._{\overline{x}}}$$

It looks just like Z; the only difference is that we calculate it from the sample and it is a small sample.

We can obtain the probability of getting certain t values similarly to the way we obtained probabilities of Z values—from an appropriate table. But it happens that while the t distribution looks like a normal Z distribution, it is just a little different, thereby giving slightly different probabilities. In fact there are many t distributions (not just one, like for Z). There is a different t distribution for each different sample size. (More will be explained about this in Section 3.10.)

In our example, where we have a mean based on 25 cases, we would need a t value of 2.06 to correspond to a probability of 0.025 (instead of the 1.96 for the Z distribution). Translating this back to the scale of sample means, if our standard error were 3.2, then the probability would be 0.025 that we would get a sample mean as large as 106.6 (which is 100 + 2.06 times 3.2) rather than 106.3 (which is 100 + 1.96 times 3.2) as in the Z distribution. This may seem like nit-picking, since the differences are so small. In fact, as the sample size approaches infinity, the t distribution becomes exactly like the Z distribution, but the differences between Z and t get larger as the sample size gets smaller, and it is always safe to use the t distribution. For example, for a mean based on five cases, the t value would be 2.78 instead of the Z of 1.96. Some t values are tabled in Appendix 1. More detailed tables are in standard statistics books.

3.10 Sample Values and Population Values Revisited

All this going back and forth between sample values and population values may be confusing. Here are the points to remember:

1. We are always interested in estimating population values from samples.
2. In some of the formulas and terms, we use population values as if we knew what the population values really are. We of course don't know the actual population values, but if we have very large samples, we can estimate them quite well from our sample data.

3. For practical purposes, we will generally use and refer to techniques appropriate for small samples, since that is more common and safer (i.e., it doesn't hurt even if we have large samples).

3.11 A Question of Confidence

A *confidence interval* establishes a range and specifies the probability that this range encompasses the true population mean. For instance, a 95% confidence interval (approximately) is set up by taking the sample mean, $\bar{x}$, plus or minus *two standard errors of the mean*.

95% confidence interval:

$$\bar{x} \pm 2\ s.e. = \bar{x} \pm 2\left(\frac{s.d.}{\sqrt{n}}\right)$$

Thus, if we took a random sample of 64 applicants to the Albert Einstein College of Medicine and found their mean IQ to be 125, say, (a fictitious figure), we might like to set up a 95% confidence interval to infer what the true mean of the population of applicants really is. The 95% confidence interval is the range between 125–2 s.e. and 125 + 2 s.e. We usually phrase this as

> We are 95% confident that the true mean IQ of Einstein medical school applicants lies within 125 ± 2 s.e.

For the purposes of this example, assume that the standard deviation is 16. (This is not a particularly good assumption since the IQ variability of medical school applicants is considerably less than the variability of IQ in the population in general.) Under this assumption, we arrive at the following range:

$$125 + \frac{2(16)}{\sqrt{64}} = 125 \pm \frac{2(16)}{8} = 125 \pm 4 = 121 - 129$$

Our statement now is as follows: "The probability is approximately 0.95 that the true mean IQ of Einstein Medical School applicants lies within the range 121–129." (A more rigorous interpretation of this is given in Section 3.11.)

A 99% confidence interval is approximately the sample mean ± 3 s.e. In our example, this interval would be

$$125 \pm 3\left[\frac{(16)}{\sqrt{64}}\right] = 125 \pm 6 = 119 - 131$$

We would then be able to say: "The probability is approximately 0.99 that the true mean IQ of Einstein Medical School applicants lies within the range 119–131."

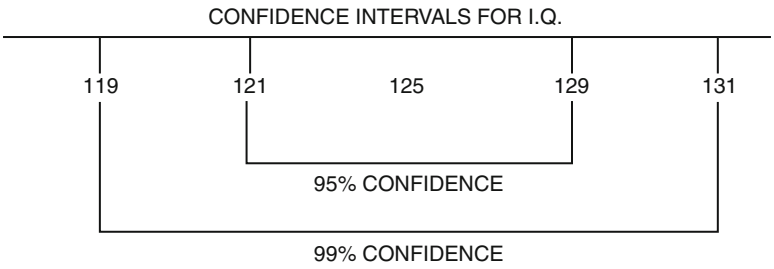


Figure 3.7 Confidence intervals

The "approximately" is because to achieve 0.95 probability, you don't multiply the s.e. by 2 exactly as we did here; we rounded it for convenience. The *exact* multiplying factor depends on how large the sample is. If the sample is very large, greater than 100, we would multiply the s.e. by 1.96 for 95% confidence intervals and by 2.58 for 99% confidence intervals. If the sample is smaller, we should look up the multiplier in tables of t values, which appear in many texts. These t values are different for different "degrees of freedom," explained in Section 3.13, which are related to sample sizes. Some t values are shown in Appendix 1. (Also refer back to Section 3.9 for the meaning of t statistics.)

Note that for a given sample size, we trade off degree of certainty for size of the interval. We can be more certain that our true mean lies within a wider range, but if we want to pin down the range more precisely, we are less certain about it (Figure 3.7). To achieve more precision and maintain a high probability of being correct in estimating the range, it is necessary to increase the sample size. The main point here is that when you report a sample mean as an estimate of a population mean, it is most desirable to report the confidence limits.

3.12 Confidence Limits and Confidence Intervals

Confidence limits are the outer boundaries that we calculate and about which we can say: We are 95% confident that these boundaries or limits include the true population mean. The interval between these limits is called the *confidence interval*. If we were to take a large number of samples from the population and calculate the 95% confidence limits for each of them, 95% of the intervals bound by these limits would contain the true population mean. However, 5% would not contain it. Of course, in real life we only take one sample and construct confidence intervals from it. We can never be sure whether the interval calculated from our particular sample is one of the 5% of such intervals that do not contain the population mean. The most we can say is that we are 95% confident it does contain it. As you can see, we never know anything for sure.

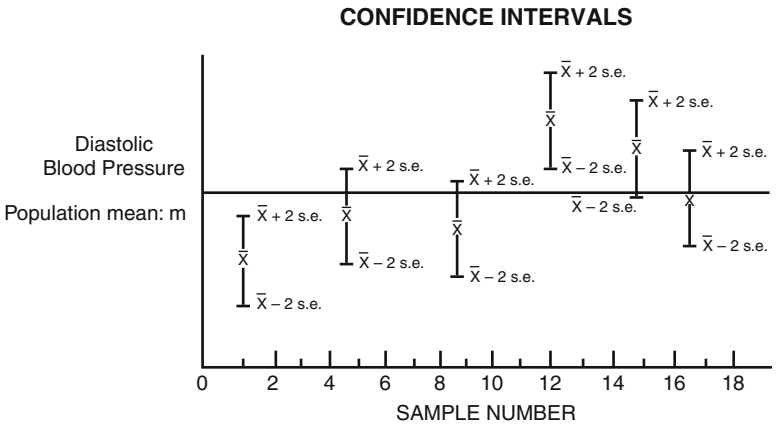


Figure 3.8 Confidence intervals

If an infinite number of independent random samples were drawn from the population of interest (with replacement), then 95% of the confidence intervals calculated from the samples (with mean $\bar{x}$ and standard error s.e.) will encompass the true population mean m.

Figure 3.8 illustrates the above concepts.

3.13 Degrees of Freedom

The t values that we use as the multiplier of the standard error to construct confidence intervals depend on something called the *degrees of freedom* (df), which are related to the sample size. When we have one sample, in order to find the appropriate t value to calculate the confidence limits, we enter the tables with n–1 degrees of freedom, where n is the sample size. An intuitive way to understand the concept of df is to consider that if we calculate the mean of a sample of, say, three values, we would have the "freedom" to vary two of them any way we liked after knowing what the mean is, but the third must be fixed in order to arrive at the given mean. So we only have two "degrees of freedom." For example, if we know the mean of three values is 7, we can have the following sets of data:

Value 1:	7	−50
Value 2:	7	+18
Value 3:	7	+53
Sum =	21	21
Mean =	$\bar{x} = 7$	$\bar{x} = 7$

In each case, if we know values 1 and 2, then value 3 is determined since the sum of these values must be 21 in order for the mean to be 7. We have "lost" one degree of freedom in calculating the mean.

3.14 Confidence Intervals for Proportions

A proportion can be considered a continuous variable. For example, in the anticoagulant study described in Section 3.1, the proportion of women in the control (placebo-treated) group who survived a heart attack was found to be 89/129 = .69. A proportion may assume values along the continuum between 0 and 1. We can construct a confidence interval around a proportion in a similar way to constructing confidence intervals around means. The 95% confidence limits for a proportion are $p \pm 1.96$ s.e.$_p$, where s.e.$_p$ is the *standard error of a proportion*.

To calculate the standard error of a proportion, we must first calculate the standard deviation of a proportion and divide it by the square root of n. We define our symbology:

$$s = standard\ \ deviation\ \ of\ \ a\ \ proportion = \sqrt{pq}$$

$$p = sample\ \ proportion = \frac{number\ \ of\ \ survivors\ \ in\ \ control\ \ group}{total\ \ number\ \ of\ \ women\ in\ \ control\ \ group}$$

$$q = 1 - p = \frac{number\ \ dead\ \ in\ \ control\ \ group}{total\ \ number\ \ of\ \ women\ \ in\ \ control\ \ group}$$

$$s.e._p = \frac{\sqrt{pq}}{\sqrt{n}} = \sqrt{\frac{pq}{n}}$$

In our example of women survivors of a heart attack in the control group, the 95% confidence interval is

$$.69 \pm 1.96x\sqrt{\frac{(.69) \times (.31)}{129}} = .69 \pm .08$$

And we can make the statement that we are 95% confident that the population proportion of untreated women surviving a heart attack is between 0.61 and 0.77 or 61% and 77%. (Remember this refers to the population from which our sample was drawn. We cannot generalize this to all women having a heart attack.)

For 99% confidence limits, we would multiply the standard error of a proportion by 2.58, to get the interval 0.59 to 0.80. The multiplier is the Z value that corresponds to 0.95 for 95% confidence limits or 0.99 probability for 99% confidence limits.

3.15 Confidence Intervals Around the Difference Between Two Means

We can construct confidence intervals around a difference between means in a similar fashion to which we constructed confidence intervals around a single mean. The 95% confidence limits around the difference between means are given by

$$(\overline{x} - \overline{y}) \pm \left(t_{df,.95}\right)\left(s.e._{\overline{x}-\overline{y}}\right)$$

In words, this is the difference between the two sample means, plus or minus an appropriate t value, times the standard error of the difference; df is the degrees of freedom and 0.95 says that we look up the t value that pertains to those degrees of freedom and to 0.95 probability. The degrees of freedom when we are looking at two samples are $n_x + n_y$–2. This is because we have lost one degree of freedom for each of the two means we have calculated, so our total degrees of freedom is $(n_x - 1) + (n_y - 1) = n_x + n_y - 2$.

As an example, consider that we have a sample of 25 female and 25 male medical students. The mean IQs for each sample are

$$\overline{x}_{females} = 130, \quad \overline{x}_{males} = 126, \quad s_{pooled} = 12, \quad df = 48$$

The 95% confidence interval for the mean difference between men and women is calculated as follows:

From t tables, we find that the t value for df = 48 is 2.01:

$$\overline{x}_{females} - \overline{x}_{males} \pm 2.01 \times \sqrt{s_p\left(\frac{1}{n_x} + \frac{1}{n_y}\right)} = (130 - 126) \pm 2.01\sqrt{12(1/25 + 1/25)}$$
$$= 4 \pm 6.8$$

The interval then is −2.8 to 10.8, and we are 95% certain it includes the true mean difference between men and women. This interval includes 0 difference, so we would have to conclude that the difference in IQ between men and women may be zero.

3.16 Comparisons Between Two Groups

A most common problem that arises is the need to compare two groups on some dimension. We may wish to determine, for instance, whether (1) administering a certain drug lowers blood pressure, or (2) drug A is more effective than drug B in lowering blood sugar levels, or (3) teaching first-grade children to read by method I produces higher reading achievement scores at the end of the year than teaching them to read by method II.

3.17 Z-Test for Comparing Two Proportions

As an example, we reproduce here the table in Section 3.1 showing data from a study on anticoagulant therapy.

Observed frequencies

	Control	Treated	
Lived	89	223	312
Died	40	39	79
Total	129	262	391

If we wish to test whether the proportion of women surviving a heart attack in the treated group differs from the proportion surviving in the control group, we set up our null hypothesis as

H_O: $P_1 = P_2$ or $P_1 - P_2 = 0$; P_1 = proportion surviving in treated population
P_2 = proportion surviving in control population
H_A: $P_1 - P_2 \neq 0$ (the difference does not equal 0)

We calculate

$$Z = \frac{p_1 - p_2}{s.e._{p_1 - p_2}}$$

$$p_1 = \frac{223}{262} = .85, \quad q_1 = 1 - p_1 = .15, \quad n_1 - 262$$

$$p_2 = \frac{89}{129} = .69, \quad q_2 = 1 - p_2 = .31, \quad n_2 = 129$$

Thus, the numerator of Z = .85–.69 = .16.
The denominator =
standard error of the difference between two proportions =

$$s.e._{(p_1 - p_2)} = \sqrt{\widehat{pq}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\widehat{p}$ and $\widehat{q}$ are pooled estimates based on both treated and control group data. We calculate it as follows:

$$\begin{aligned}
\widehat{p} &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{number\ \ of\ \ survivors\ \ in\ \ treated + control}{total\ \ number\ \ of\ \ patients\ \ in\ \ treated + control} \\
&= \frac{262(.85) + 129(.69)}{262 + 129} = \frac{223 + 89}{391} = .80 \\
\widehat{q} &= 1 - \widehat{p} = 1 - .80 = .20 \\
s.e._{(p_1 - p_2)} &= \sqrt{(.80)(.20)\left(\frac{1}{262} + \frac{1}{129}\right)} = .043 \\
Z &= \frac{.85 - .69}{.043} = 3.72
\end{aligned}$$

We must now look to see if this value of Z exceeds the critical value. The critical value is the minimum value of the test statistics that we must get in order to reject the null hypothesis at a given level of significance.

The *critical value of Z* that we need to reject H_O at the 0.05 level of significance is 1.96. The value we obtained is 3.74. This is clearly a large enough Z to reject H_O at the 0.01 level at least. The critical value for Z to reject H_O at the 0.01 level is 2.58.

Note that we came to the same conclusion using the chi-square test in Section 3.1. In fact $Z^2 = \chi^2 = (3.74)^2 = 13.99$ and the uncorrected chi-square we calculated was 13.94 (the difference is due to rounding errors). Of course the critical values of χ^2 and Z have to be looked up in their appropriate tables. Some values appear in Appendix 1.

3.18 t-Test for the Difference Between Means of Two Independent Groups: Principles

When we wanted to compare two groups on some measure that was a discrete or categorical variable, like mortality in two groups, we used the chi-square test, described in Section 3.1. Or we could use a test between proportions as described in Section 3.17. We now discuss a method of comparing two groups when the measure of interest is a continuous variable.

Let us take as an example the comparison of the ages at first pregnancy of two groups of women: those who are lawyers and those who are paralegals. Such a study might be of sociological interest, or it might be of interest to law firms, or perhaps to a baby foods company that is seeking to focus its advertising strategy more effectively.

Assuming we have taken proper samples of each group, we now have two sets of values: the ages of the lawyers (group A) and the ages of the paralegals (group B), and we have a mean age for each sample. We set up our null hypothesis as follows:

H_O: "The mean age of the population of lawyers from which we have drawn sample A is the same as the mean age of the population of paralegals from which we have drawn sample B."

Our alternate hypothesis is

H_A: "The mean ages of the two populations we have sampled are different."

In essence, then, we have drawn samples on the basis of which we will make inferences about the populations from which they came. We are subject to the same kinds of type I and type II errors we discussed before.

The general approach is as follows: We know there is variability of the scores in group A around the mean for group A and within group B around the mean for group B, simply because even within a given population, people vary. What we want to find is whether the variability between the two sample means around the grand mean of all the scores is greater than the variability of the ages within the groups around their own means. If there is as much variability within the groups as between the groups, then they probably come from the same population.

The appropriate test here is the *t*-test. We calculate a value known as t, which is equal to the difference between the two sample means divided by an appropriate standard error. The appropriate standard error is called the standard error of the difference between two means and is written as

$$s.e._{\bar{x}_1 - \bar{x}_2}$$

The distribution of t has been tabulated, and from the tables, we can obtain the probability of getting a value of t as large as the one we actually obtained under the assumption that our null hypothesis (of no difference between means) is true. If this probability is small (i.e., if it is unlikely that by chance alone we would get a value of t that large if the null hypothesis were true), we would reject the null hypothesis and accept the alternate hypothesis that there really is a difference between the means of the populations from which we have drawn the two samples.

3.19 How to Do a *t*-Test: An Example

Although t-tests can be easily performed on personal computers, an example of the calculations and interpretation is given below. This statistical test is performed to compare the means of two groups under the assumption that both samples are random and independent and come from normally distributed populations with unknown but equal variances.

$$\text{Null hypothesis}: \quad m_A = m_B, \quad \text{or the equivalent}: \quad m_A - m_B = 0$$

$$\text{Alternate hypothesis}: \quad m_A \neq m_B, \quad \text{or the equivalent}: \quad m_A - m_B \neq 0$$

[Note: When the alternate hypothesis does not specify the direction of the difference (by stating, for instance, that m_A is greater than m_B) but simply says the difference *does not equal 0*, it is called a two-tailed test. When the direction of the difference is specified, it is called a one-tailed test. More on this topic appears in Section 6.4.]

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{s_{\bar{x}_A - \bar{x}_B}}$$

Ages of sample A			Ages of sample B		
x_i	$(x_i - \bar{x}_A)$	$(x_i - \bar{x}_A)^2$	x_i	$(x_i - \bar{x}_B)$	$(x_i - \bar{x}_B)^2$
28	−3	9	24	2.4	5.76
30	−1	1	25	3.4	11.56
27	−4	16	20	−1.6	2.56
32	1	1	18	−3.6	12.96
34	3	9	21	−0.6	0.36
36	5	25	$\Sigma = 108$	$\Sigma = 0$	$\Sigma = 33.20$
30	−1	1			
$\Sigma = 217$	$\Sigma = 0$	$\Sigma = 62$			

$Mean_A = \bar{x}_A = \frac{\Sigma x_i}{n} = \frac{217}{7} = 31$; $Mean_B = \bar{x}_B = \frac{\Sigma x_i}{n} = \frac{108}{5} = 21.6$

(The subscript *i* refers to the *i*th score and is a convention used to indicate that we sum over all the scores.)

The numerator of t is the difference between the two means:

$$31 - 21.6 = 9.4$$

To get the denominator of t, we need to calculate the standard error of the difference between means, which we do as follows:

First we get the pooled estimate of the standard deviation. We calculate:

$$s_p = \sqrt{\frac{\Sigma(x_i - \bar{x}_A)^2 + \Sigma(x_i - \bar{x}_B)^2}{n_A + n_B - 2}} = \sqrt{\frac{62 + 33.20}{7 + 5 - 2}}$$

$$= \sqrt{\frac{95.20}{10}} = \sqrt{9.52} = 3.09$$

$$s_{\bar{x}_A - \bar{x}_B} = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 3.09 \sqrt{\frac{1}{7} + \frac{1}{5}} = 3.09 \sqrt{.3428} = 3.09 \times .5854 = 1.81$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_{\bar{x}_A - \bar{x}_B}} = \frac{9.4}{1.81} = 5.19$$

This t is significant at the 0.001 level, which means that you would get a value of t as high as this one or higher only one time out of a thousand by chance if the null hypothesis were true. So we reject the null hypothesis of no difference, accept the alternate hypothesis, and conclude that the lawyers are older at first pregnancy than the paralegals.

3.20 Matched-Pair t-Test

If you have a situation where the scores in one group correlate with the scores in the other group, you cannot use the regular t-test since that assumes the two groups are independent. This situation arises when you take two measures on the same individual. For instance, suppose group A represents reading scores of a group of children taken at time 1. These children have then been given special instruction in reading over a period of 6 months, and their reading achievement is again measured to see if they accomplished any gains at time 2. In such a situation, you would use a matched-pair t-test.

Child	A Initial reading scores of children	B Scores of same children after 6 months' training	d = B–A	$d - \bar{d}$	$(d - \bar{d})^2$
(1)	60	62	2	1.4	1.96
(2)	50	54	4	3.4	11.56
(3)	70	70	0	−0.6	0.36
(4)	80	78	−2	−2.6	6.76
(5)	75	74	−1	−1.6	2.56
		Sum	3	0	23.20
Mean difference = $\bar{d} = 3/5 = 0.60$					

Null hypothesis: Mean difference = 0.
Alternate hypothesis: Mean difference is greater than 0.

$$t = \frac{\bar{d}}{s_{\bar{d}}}; s_{\bar{d}} = \frac{s}{\sqrt{n}}$$

$$s = \sqrt{\frac{\Sigma (d - \bar{d})^2}{n - 1}} = \sqrt{\frac{23.20}{4}} = \sqrt{5.8} = 2.41$$

$$s_{\bar{d}} = \frac{2.41}{\sqrt{5}} = \frac{2.41}{2.23} = 1.08$$

$$t = \frac{.60}{1.08} = .56$$

This *t* is not significant, which means that we do not reject the null hypothesis and conclude that the mean difference in reading scores could be zero; that is, the 6 months' reading program may not be effective. (Or it may be that the study was not large enough to detect a difference, and we have committed a type II error.)

When the actual difference between matched pairs is not in itself a meaningful number but the researcher can *rank* the difference scores (as being larger or smaller for given pairs), the appropriate test is the Wilcoxon matched-pair rank-sum test. This is known as a *nonparametric test* and along with other such tests is described with exquisite clarity in the classic book by Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences* and in *Nonparametric Statistics for Non-Statisticians* (listed in the "Suggested Readings").

3.21 When Not to Do a Lot of t-Tests: The Problem of Multiple Tests of Significance

A *t*-test is used for comparing the means of two groups. When there are three or more group means to be compared, the *t*-test is not appropriate. To understand why, we need to invoke our knowledge of combining probabilities from Section 2.2.

Suppose you are testing the effects of three different treatments for high blood pressure. Patients in group A receive one medication, a diuretic; patients in group B receive another medication, a beta-blocker; and patients in group C receive a placebo pill. You want to know whether either drug is better than placebo in lowering blood pressure and if the two drugs are different from each other in their blood pressure-lowering effect.

There are three comparisons that can be made: group A versus group C (to see if the diuretic is better than placebo), group B versus group C (to see if the beta-blocker is better than the placebo), and group A versus group B (to see which of the two active drugs has more effect). We set our significance level at 0.05; that is, we are willing to be *wrong* in rejecting the null hypothesis of no difference between two means, with a probability of 0.05 or less (i.e., our probability of making a type I error must be no greater than 0.05). Consider the following:

Comparison	Probability of type I error	Probability of *not* making a type I error = 1–*P* (type I error)
1. A versus C	.05	1–.05 = .95
2. B versus C	.05	1–.05 = .95
3. A versus B	.05	1–.05 = .95

The probability of *not* making a type I error in the first comparison *and* not making it in the second comparison *and* not making it in the third comparison = 0.95 × 0.95 × 0.95 = 0.86. We are looking here at the *joint* occurrence of three events (the three ways of *not* committing a type I error), and we combine these probabilities by multiplying the individual probabilities. (Remember, when we

see "and" in the context of combining probabilities, we multiply, when we see "or" we add.) So now, we know that the overall probability of *not* committing a type I error in any of the three possible comparisons is 0.86. Therefore, the probability of committing such an error is 1—the probability of not committing it, or 1–0.86 = 0.14. Thus, the overall probability of a type I error would be considerably greater than the 0.05 we specified as desirable. In actual fact, the numbers are a little different because the three comparisons are not independent events, since the same groups are used in more than one comparison, so combining probabilities in this situation would not involve the simple multiplication rule for the joint occurrence of independent events. However, it is close enough to illustrate the point that making multiple comparisons in the same experiment results in quite a different significance level (0.14 in this example) than the one we chose (0.05). When there are more than three groups to compare, the situation gets worse.

3.22 Analysis of Variance: Comparison Among Several Groups

The appropriate technique for analyzing continuous variables when there are three or more groups to be compared is the analysis of variance, commonly referred to as ANOVA. An example might be comparing the blood pressure reduction effects of the three drugs.

3.23 Principles Underlying Analysis of Variance

The principles involved in the analysis of variance are the same as those in the *t*-test. Under the null hypothesis, we would have the following situation: There would be one big population and if we picked samples of a given size from that population, we would have a bunch of sample means that would vary due to chance around the grand mean of the whole population. If it turns out they vary around the grand mean more than we would expect just by chance alone, then perhaps something other than chance is operating. Perhaps they don't all come from the same population. Perhaps something distinguishes the groups we have picked. We would then reject the null hypothesis that all the means are equal and conclude the means are different from each other by more than just chance. Essentially, we want to know if the variability of all the group means is substantially greater than the variability within each of the groups around their own mean.

We calculate a quantity known as the *between-groups variance*, which is the variability of the group means around the grand mean of all the data. We calculate another quantity called the *within-groups variance*, which is the variability of the scores within each group around its own mean. One of the assumptions of the

analysis of variance is that the extent of the variability of individuals within groups is the same for each of the groups, so we can pool the estimates of the individual within-group variances to obtain a more reliable estimate of overall within-groups variance. If there is as much variability of individuals *within* the groups as there is variability of means *between* the groups, the means probably come from the same population, which would be consistent with the hypothesis of no true difference among means; that is, we could not reject the null hypothesis of no difference among means.

The ratio of the between-groups variance to the within-groups variance is known as the F ratio. Values of the F distribution appear in tables in many statistical texts, and if the obtained value from our experiment is greater than the *critical value* that is tabled, we can then reject the hypothesis of no difference.

There are different critical values of F, depending on how many groups are compared and on how many scores there are in each group. To read the tables of F, one must know the two values of degrees of freedom (df). The df corresponding to the between-groups variance, which is the numerator of the F ratio, is equal to k–1, where k is the number of groups. The df corresponding to the denominator of the F ratio, which is the within-groups variance, is equal to $k \times (n–1)$, that is, the number of groups times the number of scores in each group minus one. For example, if in our hypertension experiment there are 100 patients in each of the 3 drug groups, then the numerator degrees of freedom would be $3–1 = 2$, and the denominator degrees of freedom would be $3 \times 99 = 297$. An F ratio would have to be at least 3.02 for a significance level of 0.05. If there were four groups being compared, then the numerator degrees of freedom would be 3, and the critical value of F would need to be 2.63. If there is not an equal number of individuals in each group, then the denominator degrees of freedom is $(n_1–1) + (n_2–1) + (n_3–1)$.

We will not present here the actual calculations necessary to do an F test because nowadays these are rarely done by hand. There are a large number of programs available for personal computers that can perform F tests, t-tests, and most other statistical analyses. However, shown below is the kind of output that can be expected from these programs. Shown are summary data from the Trial of Antihypertensive Interventions and Management (TAIM) study. The TAIM study was designed to evaluate the effect of diet and drugs, used alone or in combination with each other, to treat overweight persons with mild hypertension (high blood pressure)[10, 11].

The next table shows the mean drop in blood pressure after 6 months of treatment with each drug, the number of people in each group, and the standard deviation of the *change* in blood pressure in each group.

Drug group	*n*	Mean drop (in diastolic blood pressure units after 6 months of treatment)	Standard deviation
A. Diuretic	261	12.1	7.9
B. Beta-blocker	264	13.5	8.2
C. Placebo	257	9.8	8.3

The next table results from an analysis of variance of the data from this study. It is to be interpreted as follows:

ANOVA					
Source of variation	Degrees of freedom	Sum of squares	Mean square	F ratio	$P_2 > F$
Between groups	2	1776.5	888.2	13.42	.0001
Within groups	779	5256.9	66.2		
	781				

The mean square is the sum of squares divided by the degrees of freedom. For between-groups, it is the variation of the group means around the grand mean, while for within-groups, it is the pooled estimate of the variation of the individual scores around their respective group means. The within-groups mean square is also called the error mean square. (An important point is that the square root of the error mean square is the pooled estimate of the within-groups standard deviation. In this case, it is $\sqrt{66.2} = 8.14$. It is roughly equivalent to the average standard deviation.) F is the ratio of the between to the within mean squares; in this example, it is $888.2/66.2 = 13.42$.

The F ratio is significant at the 0.0001 level, so we can reject the null hypothesis that *all* group means are equal. However, we do not know where the difference lies. Is group A different from group C but not from group B? We should not simply make all the pairwise comparisons possible because of the problem of multiple comparisons discussed above. But there are ways to handle this problem. One of them is the Bonferroni procedure, described in the next section.

3.24 Bonferroni Procedure: An Approach to Making Multiple Comparisons

This is one way to handle the problem of multiple comparisons. The Bonferroni procedure implies that if, for example, we make five comparisons, the probability that *none* of the five p values falls below 0.05 is at least $1–(5 \times 0.05) = 0.75$ when the null hypothesis of equal means is really true. That means that there is a probability of up to 0.25 that at least one p value will reach the 0.05 significance level by chance alone *even if the treatments really do not differ*. To get around this, we divide the chosen overall significance level by the number of two-way comparisons to be made, consider this value to be the significance level for any single comparison, and reject the null hypothesis of no difference only if it achieves this new significance level.

For example, if we want an overall significance level of 0.05 and we will make three comparisons between means, we would have to achieve $0.05/3 = 0.017$ level in order to reject the null hypothesis and conclude there is a difference between the two means. A good deal of self-discipline is required to stick to this procedure and not declare a difference between two means as unlikely to be due to chance if the particular comparison has significance at $p = .03$, say, instead of 0.017. The

Bonferroni procedure does not require a prior F test. Let us apply the Bonferroni procedure to our data.

First we compare each of the drugs to placebo. We calculate the t for the difference between means of group A versus group C.

$$t = \frac{\bar{x}_A - \bar{x}_C}{s.e._{\bar{x}_A - \bar{x}_C}}$$

$$s.e._{\bar{x}_A - \bar{x}_C} = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_C}}$$

$$\frac{12.1 - 9.8}{8.14\sqrt{\frac{1}{261} + \frac{1}{257}}} = \frac{2.3}{.715} = 3.22$$

$$p = .0014$$

Note that we use 8.14 as *s* pooled. We obtained this from the analysis of variance as an estimate of the common standard deviation. The degrees of freedom to enter the t tables are 261 + 257–2 = 516.

It turns out that the probability of getting such a high t value by chance is only 0.0014. We can safely say the diuretic reduces blood pressure more than the placebo. The same holds true for the comparison between the beta-blocker and placebo. Now let us compare the two drugs, B versus A:

$$t = \frac{13.5 - 12.1}{8.14\sqrt{\frac{1}{264} + \frac{1}{261}}} = \frac{1.4}{.711} = 1.97$$

The *p* value corresponding to this t value is 0.049. It might be tempting to declare a significant difference at the 0.05 level, but remember the Bonferroni procedure requires that we get a *p* value of 0.017 or less for significance adjusted for multiple comparisons. The critical value of t corresponding to $p = 0.017$ is 2.39, and we only got a t of 1.97. However, there has been some questioning of the routine adjustment for multiple comparisons[12] on the grounds that we thereby may commit more type II errors and miss important effects. In any case, *p* levels should be reported so that the informed reader may evaluate the evidence.

3.25 Analysis of Variance When There Are Two Independent Variables: The Two-Factor ANOVA

The example above is referred to as the one-way ANOVA because you can divide all the scores in one way only, by the drug group to which patients were assigned. The drug group is called a "factor," and this factor has three levels, meaning there are three categories of drug. There may, however, be another factor that classifies

individuals, and in that case, we would have a two-way, or a two-factor, ANOVA. In the experiment we used as an example, patients were assigned to one of the three drugs noted above as well as to one of three diet regimens—weight reduction, sodium (salt) restriction, or no change from their usual diet, which is analogous to a placebo diet condition. The diagram below illustrates this two-factor design, and the mean drop in blood pressure in each group, as well as the numbers of cases in each group, which are shown in parenthesis.

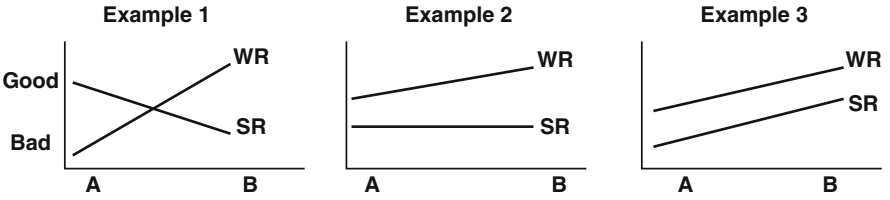
Now we are interested in comparing the three means that represent change in blood pressure in the drug groups, the three means that represent changes in the diet groups, and the interaction between drug and diet. We now explain the concept of interaction.

	Diet			
Drug	Usual	Weight reduction	Sodium restriction	Total
Diuretic	10.2 (87)	14.5 (86)	11.6 (88)	12.1 (261)
Beta-blocker	12.8 (86)	15.2 (88)	12.6 (90)	13.5 (264)
Placebo	8.7 (89)	10.8 (89)	10.1 (79)	9.8 (257)
Total	10.5 (262)	13.5 (263)	11.5 (257)	

3.26 Interaction Between Two Independent Variables

Interaction between two independent variables refers to differences in the effect of one variable, depending on the level of the second variable. For example, maybe one drug produces better effects when combined with a weight-reduction diet than when combined with a sodium-restricted diet. There may not be a significant effect of that drug when all diet groups are lumped together, but if we look at the effects separately for each diet group, we may discover an interaction between the two factors: diet and drug.

The diagrams below illustrate the concept of interaction effects. WR means weight reduction and SR means sodium (salt) restriction.



In example 1, drug B is better than drug A in those under weight reduction, but in those under salt restriction, drug A is better than drug B. If we just compared the average for drug A, combining diets, with the average for drug B, we would have to say there is no difference between drug A and drug B, but if we look at the two diets separately, we see quite different effects of the two drugs.

In example 2, there is no difference in the two drugs for those who restrict salt, but there is less effect of drug A than drug B for those in weight reduction.

In example 3, there is no interaction; there is an equal effect for both diets: The two lines are parallel; their slopes are the same. Drug B is better than drug A both for those in weight reduction and salt restriction.

3.27 Example of a Two-Way ANOVA

Next is a table of data from the TAIM study showing the results of a *two-way analysis of variance*:

Two-way ANOVA from TAIM data

Source	DF	ANOVA sum of squares	Mean square	F value	Probability
Drug group	2	1776.49	888.25	13.68	.0001
Diet group	2	1165.93	582.96	8.98	.0001
Drug × diet	4	214.50	53.63	0.83	.509
Error	773	50,185.46	64.93		

Note that the error mean square here is 64.93 instead of 66.9 when we did the one-way analysis. That is because we have explained some of the error variance as being due to diet effects and interaction effects (we have "taken out" these effects from the error variance). Thus, 64.93 represents the variance due to pure error, or the unexplained variance. Now we can use the square root of this, which is 8.06 as the estimate of the common standard deviation. We explain the results as follows: There is a significant effect of drug ($p = .0001$) and a significant effect of diet ($p = 0.0001$), but no interaction of drug by diet ($p = .509$).

We have already made the three pairwise comparisons, by t-tests for the difference between two means among drugs (i.e., placebo vs. diuretic, placebo vs. beta-blocker, and diuretic vs. beta-blocker). We can do the same for the three diets. Their mean values are displayed below:

Diet group	n	Mean drop in diastolic blood pressure	Standard deviation
Weight reduction	263	13.5	8.3
Sodium restriction	257	11.5	8.3
Usual diet	262	10.5	8.0

Pooled estimate of s.d. = 8.06

If we did t-tests, we would find that weight reduction is better than usual diet ($p = .0000$), but sodium restriction shows no significant improvement over usual diet ($p = .16$).

Weight reduction when compared with sodium restriction is also significantly better with $p = .005$, which is well below the $p = .017$ required by the Bonferroni procedure. (The t for this pairwise comparison is 2.83, which is above the critical value of 2.39.)

3.28 Kruskal–Wallis Test to Compare Several Groups

The analysis of variance is valid when these conditions are met: 1) the variable of interest is continuous and is normally distributed (i.e. it has the familar bell-shaped curve); 2) the individual variances within each of the groups being compared are essentially very similar. Often, however, we must deal with situations when we want to compare several groups on a variable that does not meet all of the above conditions. This might be a case where we can say one person is better than another, but we can't say exactly how much better. In such a case, we would rank people and compare the groups by using the Kruskal–Wallis test to determine if it is likely that all the groups come from a common population. This test is analogous to the one-way analysis of variance, but instead of using the original scores, it uses the *rankings of* the scores. It is called a *nonparametric test*. This test is available in many computer programs, but an example appears in Appendix 3.

3.29 Association and Causation: The Correlation Coefficient

A common class of problems in the accumulation and evaluation of scientific evidence is the assessment of association of two variables. Is there an association between poverty and drug addiction? Is emotional stress associated with cardiovascular disease?

To determine association, we must first quantify both variables. For instance, emotional stress may be quantified by using an appropriate psychological test of stress or by clearly defining, evaluating, and rating on a scale the stress factor in an individual's life situation, whereas hypertension (defined as a blood pressure

reading) may be considered as the particular aspect of cardiovascular disease to be studied. When variables have been quantified, a measure of association needs to be calculated to determine the strength of the relationship. One of the most common measures of association is the *correlation coefficient*, r, which is a number derived from the data and which can vary between -1 and $+1$. (Most statistical computer packages will calculate the correlation coefficient, but if you want to do it yourself, the method of calculation appears in Appendix 4.) When $r = 0$, it means there is no association between the two variables. An example of this might be the correlation between blood pressure and the number of hairs on the head. When $r = +1$, a perfect positive correlation, it means there is a direct relationship between the two variables: An individual who has a high score on one variable also has a high score on the other, and the score on one variable can be exactly predicted from the score on the other variable. This kind of correlation exists only in deterministic models, where there is really a functional relationship. An example might be the correlation between age of a tree and the number of rings it has. A correlation coefficient of -1 indicates a perfect inverse relationship, where a high score on one variable means a low score on the other and where, as in perfect positive correlation, there is no error of measurement. Correlation coefficients between 0 and $+1$ and between 0 and -1 indicate varying strengths of associations.

These correlation coefficients apply when the basic relationship between the two variables is linear. Consider a group of people for each of whom we have a measurement of weight against height; we will find that we can draw a straight line through the points. There is a linear association between weight and height, and the correlation coefficient would be positive but less than 1. When the variables are continuous, the calculated correlation is the *Pearson product–moment correlation.* If the variables are ranked and ordered according to rank, we calculate the *Spearman rank-order correlation*, which is a nonparametric statistic. Nonparametric statistics are used when the data do not have to be normally distributed and are ordinal (i.e., can be sorted in order, but the distances between any two values do not have to be the same). An example is educational level, which can be categorized into less than a high school education, graduated from high school, some college, graduated from college, and received a graduate degree. You can assign numbers to these from 1 to 5, but the numbers do not represent years of education but rather categories of education that are ordered from lowest to highest category, and you can categorize them differently if you wish.

The diagrams in Figure 3.9 illustrate representations of various correlation coefficients.

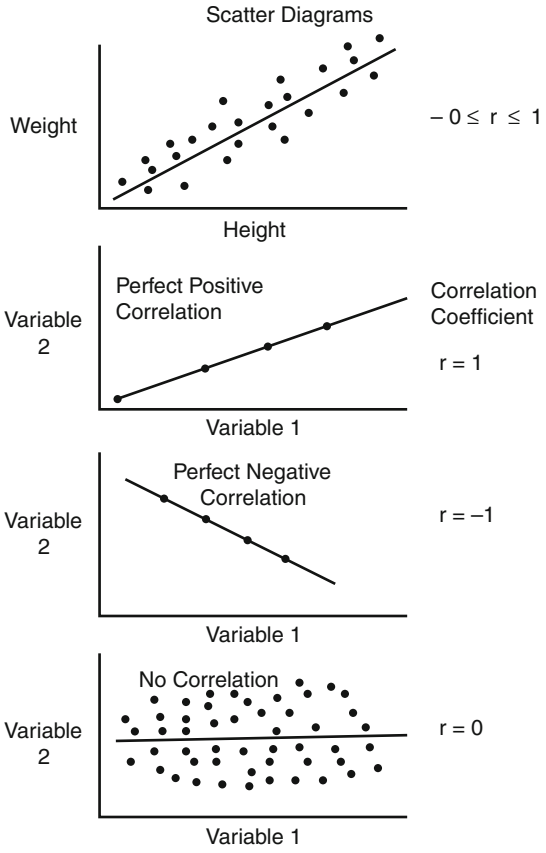


Figure 3.9 Correlations between two variables

3.30 Some Points to Remember About Correlation

- A correlation coefficient squared (r^2) tells you what proportion of the variance in one variable is explained by the other variable. Thus, $r = .40$ means that $0.4 \times 0.4 = 0.16$ or 16% of the variation in one variable is explained by the other variable, but the remaining 84% of the variation is due to other factors.
- How high is high? The answer to this question depends upon the field of application as well as on many other factors, including precision of measurement, as well as the fact that there may be different relationships between two variables in different groups of people. For example, the correlations between verbal aptitude and nonverbal aptitude, as measured for Philadelphia schoolchildren by standardized national tests, range from 0.44 to 0.71 depending on race and social class of the groups.[13]
- Size of the correlation coefficient and statistical significance. A very low correlation coefficient may be statistically significant if the sample size is large enough. What seems to be a high correlation coefficient may not reach statistical

significance if the sample size is too low. Statistical significance tells you that the correlation you observed is not likely due to chance, but does not tell you anything about the strength of the association.

- Correlation is not causation. This is the most important point and one that is often ignored. Correlation tells you nothing about the directionality of the relationship, nor about whether there is some third factor influencing both variables. The most famous case is that which is invoked to support the theory that storks bring babies. There is a high correlation between the number of strokes in an area and the number of babies born. This was observed in the city of Copenhagen by looking at records of number of births and the number of storks over a 10-year period after World War II.[1] It turns out that after the war, more people moved to Copenhagen (so more babies were born there) and more construction was going on in the city (so storks had more nesting places, thus more storks). An amusing parody on this topic is by Thomas Hofer.[2]

3.31 Causal Pathways Underlying Correlations

If we do get a significant correlation, we then ask what situations could be responsible for it. Figure 3.10 illustrates some possible structural relationships that may underlie a significant correlation coefficient.[14] We consider two variables, W (weight gain) and B (blood pressure), and let r_{WB} represent the correlation between them. Note that only in diagrams (1), (2), and (6) do the correlation between W and B arise due to a causal relationship between the two variables. In diagram (1), W entirely determines B; in diagram (2), W is a partial cause of B; in diagram (6), W is one of several determinants of B. In all of the other structural relationships, the correlation between W and B arises due to common influences on both variables. Thus, it must be stressed that *the existence of a correlation between two variables does not necessarily imply causation*. Correlations may arise because one variable is the partial cause of another or the two correlated variables have a common cause. Other factors, such as sampling, the variation in the two populations, and so on, affect the size of the correlation coefficient also. Thus, care must be taken in interpreting these coefficients.

[1] Reference thanks to Professor Angela Pignotti of Modesta Junior College, Modesta California, Ornithologische Monatsberichte, 44 No. 2, Jahrgang, 1936, Berlin Ornithologische Monatsberichte, 48 No. 1, Jahrgang, 1940, Berlin Statistiches Jahrbuch Deutscher Gemeinden, 27–33, Jahrgang, 1932–1938, Gustav Fischer, Jena.

[2] New evidence for the Theory of the Stork, Thomas Höfera, Hildegard Przyrembelb and Silvia Verlegerc, Federal Institute for Risk Assessment, Berlin, Office of the National Breast Feeding Committee at BfR, Berlin, and Independent Midwife, Berlin, Germany, Blackwell Publishing Ltd. Paediatric and Perinatal Epidemiology 2004, 18, 88–92

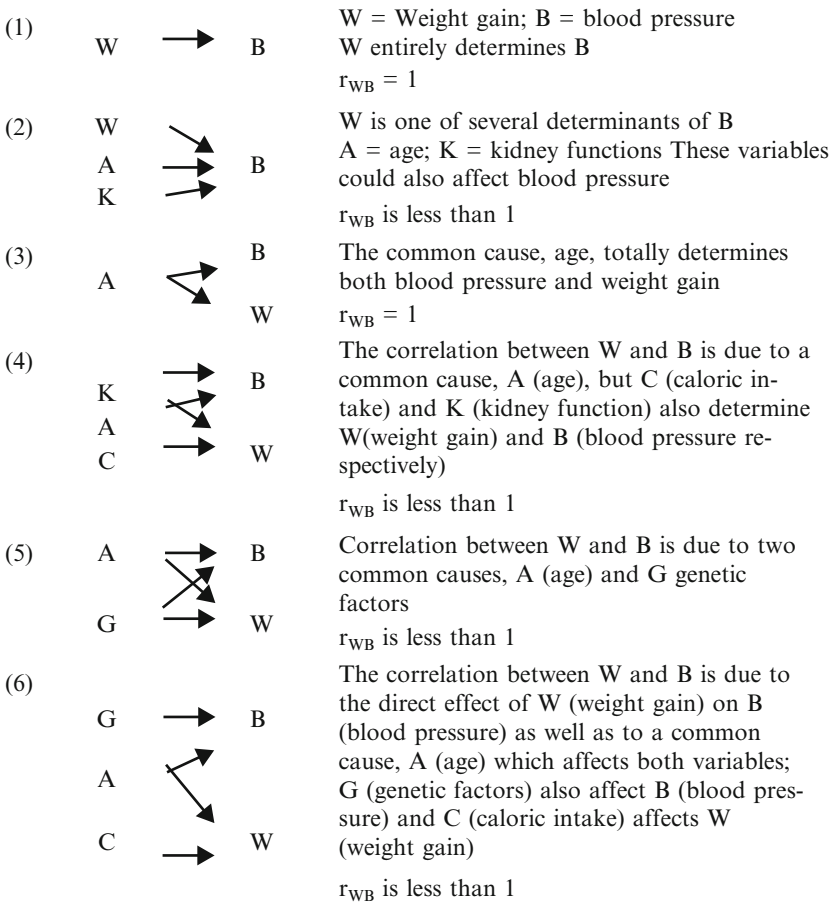


Figure 3.10 Possible relationships that underlie correlations

3.32 Regression

Note that in Figure 3.9 we have drawn lines that seem to best fit the data points. These are called *regression lines.* They have the following form $Y = a + bX$. In the top scattergram-labeled (a), *Y* is the dependent variable weight and *X*, or height, is the independent variable. We say that weight is a function of height. The quantity *a* is the intercept. It is where the line crosses the *Y* axis. The quantity *b* is the slope and it is the rate of change in *Y* for a unit change in *X*. If the slope is 0, it means we have a straight line parallel to the *x* axis, as in the illustration (d). It also means that we cannot predict *Y* from a knowledge of *X* since there is no relationship between *Y* and *X*. If we have the situation shown in scattergrams (b) or (c), we know exactly how *Y* changes when *X* changes and we can perfectly predict *Y* from a knowledge of

X with no error. In the scattergram (a), we can see that as *X* increases, *Y* increases, but we can't predict *Y* perfectly because the points are scattered around the line we have drawn. We can, however, put confidence limits around our prediction, but first we must determine the form of the line we should draw through the points. We must estimate the values for the intercept and slope. This is done by finding the "best-fit line."

The line that fits the points best has the following characteristics: If we take each of the data points and calculate its vertical distance from the line and then square that distance, the sum of those squared distances will be smaller than the sum of such squared distances from any other line we might draw. This is called the *least-squares* fit. Consider the data below where *Y* could be a score on one test and *X* could be a score on another test.

	Score	
Individual	*X*	*Y*
A	5	7
B	8	4
C	15	8
D	20	10
E	25	14

The calculations to determine the best-fit line are shown in Appendix 4. However, most statistical computer packages for personal computers provide a linear regression program that does these calculations. Figure 3.11 illustrates these points plotted in a scattergram and shows the least-squares line.

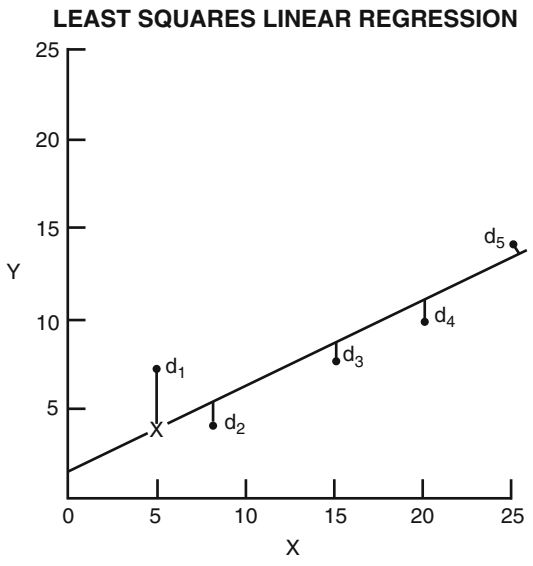


Figure 3.11 Least squares regression line

The equation for the line is $Y = 2.76 + .40\,X$. The intercept a is 2.76 so that the line crosses the y axis at $Y = 2.76$. The slope is 0.40. For example, we can calculate a predicted Y for $X = 10$ to get

$$Y = 2.76 + (.40)(10) = 2.76 + 4 = 6.76$$

The d_is are distances from the points to the line. It is the sum of these squared distances that is smaller for this line than it would be for any other line we might draw.

The correlation coefficient for these data is 0.89. The square of the correlation coefficient, r^2, can be interpreted as the proportion of the variance in Y that is explained by X. In our example, $0.89^2 = 0.79$; thus 79% of the variation of Y is explainable by the variable X, and 21% is unaccounted for.

3.33 The Connection Between Linear Regression and the Correlation Coefficient

The correlation coefficient and the slope of the linear regression line are related by the formulas

$$r = b\frac{s_x}{s_y}, \quad b = r\frac{s_y}{s_x}$$

where s_x is the standard deviation of the X variable, s_y is the standard deviation of the Y variable, b is the slope of the line, and r is the correlation coefficient.

3.34 Multiple Linear Regression

When we have two or more independent variables and a continuous dependent variable, we can use multiple regression analysis. The form this takes is

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_xX_k$$

For example, Y may be blood pressure and X_1 may be age, X_2 may be weight, and X_3 may be family history of high blood pressure. We can have as many variables as appropriate, where the last variable is the kth variable. The b_is are *regression coefficients*. Note that family history of high blood pressure is not a continuous variable. It can either be yes or no. We call this a dichotomous variable, and we can use it as any other variable in a regression equation by assigning a number to each of the two possible answers, usually by making a yes answer = 1 and a no answer = 0.

An example from the TAIM study follows and is meant only to give you an idea of how to interpret a multiple regression equation. This analysis pertains to the group of 89 people who were assigned to a placebo drug and a weight-reduction regimen. The dependent variable is change in blood pressure.

The independent variables are shown below:

Variable	Coefficient: b_i	p
Intercept	−15.49	.0016
Age	.077	.359
Race 1 = black 0 = nonblack	4.22	.021
Sex 1 = male 0 = female	1.50	.390
Pounds lost	.13	.003

Note: Sex is coded as 1 if male and 0 if female; race is coded as 1 if black and 0 if nonblack; p is used to test if the coefficient is significantly different from 0. The equation, then, is

$$\text{change in blood pressure} = -15.49 + .077_{(\text{age})} + 4.22_{(\text{race})} + 1.5_{(\text{sex})} + .13_{(\text{Change in weight})}$$

Age is not significant ($p = .359$) nor is sex ($p = .390$). However, race is significant ($p = .021$), indicating that blacks were more likely than nonblacks to have a drop in blood pressure while simultaneously controlling for all the other variables in the equation. Pounds lost are also significant, indicating that the greater the weight loss, the greater was the drop in blood pressure.

Linear regression models assume that the observations are independent of each other, i.e., they are uncorrelated. In the example above, change in weight for another person is independent of change in weight for another person in the sample. But in longitudinal studies, where individuals have repeated measures, the observations are correlated because they are observations on the same persons. Here it is appropriate to use a technique for regressions on correlated data called *GEE*, standing for generalized estimating equations, described in an article by Hanley and colleagues.[15]

3.35 Fixed Effects, Random Effects, and Mixed Models in Regression

Consider a regression model such as $Y = \beta 1X1 + \beta 2X2 + \varepsilon$ where X1 is the predictor variable of interest (known as VOI) and is categorical, X_2 is another variable, and ε is the error term.

A ***fixed effect*** is the effect of a variable whose levels or categories are specifically the ones you are interested in. For example, if you are specifically interested in

comparing males to females, then sex, which is coded in mutually exclusive categories as male or female, is a *fixed effects* variable.

A ***random effects*** variable is one whose levels are a sample from a population of such levels; they are not the specific units you are interested in, but rather they come from a population of such units that you happen to sample.

Suppose you have a study of a new surgical technique. The outcome is days till hospital discharge. Your hypothesis is that the outcome will be more favorable in women than men. Thus, sex is your VOI. The study is taking place in several hospitals around the country. You want to include the variable "hospital," which we will call X_2, in your model to control for differences among hospitals, but you are not interested in the specific hospitals. Rather, the hospitals in which the study is taking place are a sample from the population of hospitals in the United States. Thus, X_2 is a *random effects* variable.

Note that observations in the random effects variable "hospitals" are correlated. For example, the surgeons in one hospital may be better trained than in another so the outcomes for patients in one hospital would be more alike each other than they are similar to those in another hospital. Ordinary regression models assume the observations are independent, but here is an example where they are not independent but cluster in a hospital. In such a case, the average days till discharge in one hospital may differ from that in another, and each hospital has its own variance. If you were to draw a line representing the days to discharge on the Y axis and sex on the x axis, you might find that the points in hospital 1 have a different intercept than the points in hospital 2. They may also have different slopes.

A model that has both fixed and random effects is called a *mixed model*. Most computer statistical programs, like SAS, allow you to specify which variables are fixed and which are random. If you don't specify, the default is fixed effects. In practice this makes little difference in your point estimate of the size of the effect of that variable on the outcome, but it does change the variance of your estimate and so may affect significance levels. In treating a variable as a random effects variable, you separate out the variance in the residuals from the overall regression line, into two parts: the part coming from the variance due to clusters and the part due to the error variance. Thus, generally your error variance is smaller because you have taken out the component due to clusters, and your test statistic, which is based on the parameter estimate divided by the square root of the error variance, is larger and thus more likely to be significant.

In the mixed effects model, we have:

$$\begin{aligned} Y\ (\text{outcome}) = \alpha\ (\text{intercept}) + \beta_1 X_1\ (\text{fixed effect}) + \beta_2 X_2\ (\text{random effect}) \\ + \gamma\ (\text{random effect variance}) + \varepsilon\ (\text{residual variance}). \end{aligned}$$

There are several complications in the assumptions and caveats in using mixed models that are beyond the scope of this book and require more advanced mathematics. The reader is referred to other sources provided in the "References" section.

3.36 Poisson Model

When we have discussed regression models, we have assumed that the variables in the models are normally distributed. Sometimes however, we are interested in a variable that is a count variable. For example, we might consider a staffing problem for a mental health hotline. We want to be sure we have enough staff to answer the phones so that clients don't have too long a wait, but at the same time, we have budget constraints and don't want to have staff sitting around when not needed. We have basic staffing always available to handle up to four calls in an hour, but we want to know whether we have to add staff in the hours between 5 am and 7 am, which may be peak hours. Thus, we are interested in the probability that there will be more than four calls during those early morning hours. We decide that if that probability is more than 10%, we will hire another staff member between 5 and 7 am.

We can get that probability from a Poisson model, which is appropriate when (1) the variable we are interested in is a count of things, (2) the observations are independent of each other (i.e., a call received at the crisis center at 6 am is independent of whether there was a call received at 5 am), and (3) the average count in a fixed period is small.

Let us assume that the average number of calls over a 24-h period is three calls per hour. This is called λ, the mean of the Poisson distribution. So we want to know what is the probability that there will be *more than four calls* an hour when the average number of calls is 3.

Let X be the number of calls and Y is the probability of a specific number (k) of calls.

The probability we want is:

$$\mathrm{P}(\mathrm{X} = 5 \text{ or more calls}) = 1 - \mathrm{P}(\text{ 4 or fewer calls}) =$$

$$\mathrm{P}(\mathrm{k} < = 4 \text{ calls}) = 1 - \mathrm{P}(\mathrm{k} = 0 \text{ or } \mathrm{k} = 1 \text{ or } \mathrm{k} = 2 \text{ or } \mathrm{k} = 3 \text{ or } \mathrm{k} = 4)$$

$$\text{Poisson probabilty of } \mathrm{X} = \mathrm{k} \text{ is } \frac{e^{-\lambda}\ \lambda^{k}}{k!}$$

Note that $4! = 4 \times 3 \times 2 \times 1$; $3! = 3 \times 2 \times 1$; $2! = 2 \times 1$; $1! = 1$; and $0! = 1$.

$$\begin{aligned} P(5 \textit{ or more}) &= 1 - \left(\frac{e^{-}\ \lambda^{0}}{0!} + \frac{e^{-}\ \lambda^{1}}{1!} + \frac{e^{-}\ \lambda^{2}}{2!} + \frac{e^{-}\ \lambda^{3}}{3!} + \frac{e^{-}\ \lambda^{4}}{4!} \right) \\ &= 1 - \left(\frac{e^{-3}\ 3^{0}}{0!} + \frac{e^{-3}\ 3^{1}}{1!} + \frac{e^{-3}\ 3^{2}}{2!} + \frac{e^{-3}\ 3^{3}}{3!} + \frac{e^{-3}\ 3^{4}}{4!} \right) \\ &= 1 - (0.05 + 0.15 + 0.22 + 0.22 + 0.17\ ) = 1\ 0.82 = 0.18 \text{ or } 18\% \end{aligned}$$

Since this is more than our 10% threshold of probability that there will be five or more calls, we will hire another staff member between 5 and 7 AM.

3.37 Poisson Regression

The Poisson regression model is given below, where the explanatory variables are denoted by X_i. For example there may be more calls to the hotline on rainy days or in winter, so X_1 could be rain versus no rain; X_2 could be season = winter compared to summer; X_3 could be spring or fall compared to summer:

$$\lambda = e^{(\alpha+\beta_1 X_1 +\ldots+\beta_k X_k)}$$

$$\ln\ \lambda = (\alpha + \beta_1 X_1 + \ldots + \beta_k X_k)$$

log(number of calls) = Intercept + β_1(rain=1) + β_2(winter = 1) + β_3(spring or fall =1)

Thus, number of calls = $e^{(\text{Intercept} + \beta_1 (\text{rain}=1) + \beta_2 (\text{winter} = 1) + \beta_3 (\text{spring or fall} =1)}$
$= e^{(\text{Intercept})} * e^{\beta_1 (\text{rain}=1)} * e^{\beta_2 (\text{winter} = 1)} * e^{\beta_3 (\text{spring or fall} =1)}$

In a fictitious dataset using SAS program, we found, as below, that β_1 was 0.7075, significant at p= 0.004

We calculate $e^{\beta_1} = 2.03$ and we can estimate that when it rains there will be twice as many calls as when it does not rain, controlling for season.

Parameter	df	Estimate β	Standard error	Pr > ChiSq
Intercept	1	0.5308	0.3220	0.0993
Rain	1	0.7075	0.2481	0.0044
Winter	1	0.6718	0.3344	0.0446
Spring_or_fall	1	0.3798	0.3538	0.2830

3.38 Summary So Far

Investigation of a scientific issue often requires statistical analysis, especially where there is variability with respect to the characteristics of interest. The variability may arise from two sources: the characteristic may be inherently variable in the population and/or there may be error of measurement.

In this chapter, we have pointed out that in order to evaluate a program or a drug, to compare two groups on some characteristic, and to conduct a scientific investigation of any issue, it is necessary to quantify the variables.

Variables may be quantified as discrete or as continuous, and there are appropriate statistical techniques that deal with each of these. We have considered here the chi-square test, confidence intervals, Z-test, *t*-test, analysis of variance, correlation, and regression. We have pointed out that in hypothesis testing, we are subject to two kinds of errors: the error of rejecting a hypothesis when in fact it is true and the error

of accepting a hypothesis when in fact it is false. The aim of a well-designed study is to minimize the probability of making these types of errors. Statistics will not substitute for good experimental design, but it is a necessary tool to evaluate scientific evidence obtained from well-designed studies.

Philosophically speaking, statistics is a reflection of life in two important respects: (1) As in life, we can never be certain of anything (but in statistics we can calculate a probability figure describing the degree of our uncertainty), and (2) all is a trade-off—in statistics, between certainty and precision or between two kinds of error; in life, well, fill in your own trade-offs.

Chapter 4
Mostly About Epidemiology

Medicine to produce health has to examine disease; and music to create harmony must investigate discord.

Plutarch (A.D. 46–120).

4.1 The Uses of Epidemiology

Epidemiology may be defined as the study of the distribution of health and disease in *groups of people* and the study of the factors that influence this distribution. Modern epidemiology also encompasses the evaluation of diagnostic and therapeutic modalities and the delivery of health-care services. There is a progression in the scientific process (along the dimension of increasing credibility of evidence), from casual observation to hypothesis formation, to controlled observation, and to experimental studies. Figure 4.1 is a schematic representation of the uses of epidemiology. The tools used in this endeavor are in the province of epidemiology and biostatistics. The techniques used in these disciplines enable "medical detectives" to uncover a medical problem, to evaluate the evidence about its causality or etiology, and to evaluate therapeutic interventions to combat the problem.

Descriptive epidemiology provides information on the pattern of diseases, on "who has what and how much of it," information that is essential for health-care planning and rational allocation of resources. Such information may often uncover patterns of occurrence suggesting etiologic relationships and can lead to preventive strategies. Such studies are usually of the cross-sectional type and lead to the formation of hypotheses that can then be tested in case–control, prospective, and experimental studies. Clinical trials and other types of controlled studies serve to evaluate therapeutic modalities and other means of interventions and thus ultimately determine standards of medical practice, which in turn have impact on health-care planning decisions. In the following section, we will consider selected epidemiologic concepts.

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_4

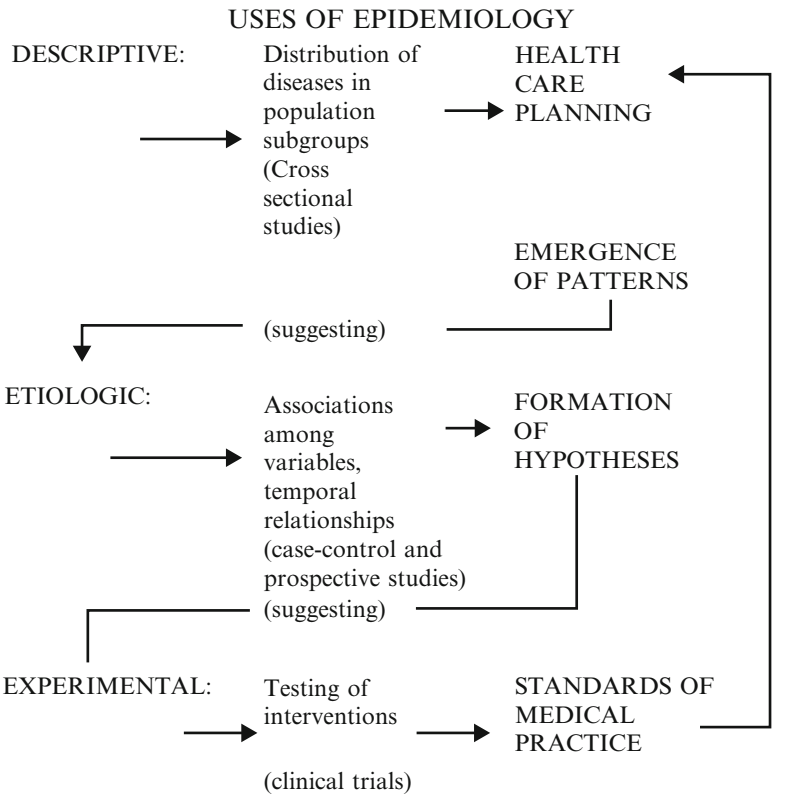


Figure 4.1 Uses of epidemiology

4.2 Some Epidemiologic Concepts: Mortality Rates

In 1900, the three major causes of death were influenza or pneumonia, tuberculosis, and gastroenteritis. Today, the three major causes of death are heart disease, cancer, and accidents; the fourth is strokes. Stroke deaths have decreased dramatically over the last few decades probably due to the improved control of hypertension, one of the primary risk factors for stroke. These changing patterns of mortality reflect changing environmental conditions, a shift from acute to chronic illness, and an aging population subject to degenerative diseases. We know this from an analysis of *rates*.

The comparison of defined rates among different subgroups of individuals may yield clues to the existence of a health problem and may lead to the specification of conditions under which this identified health problem is likely to appear and flourish.

In using rates, the following points must be remembered:

(1) A rate is a proportion involving a numerator and a denominator.
(2) Both the numerator and the denominator must be clearly defined so that you know to which group (denominator) your rate refers.

(3) The numerator is contained in (is a subset of) the denominator. This is in contrast to a ratio where the numerator refers to a different group from the denominator.

Mortality rates pertain to the number of deaths occurring in a particular population subgroup and often provide one of the first indications of a health problem. The following definitions are necessary before we continue our discussion:

The crude annual mortality rate (or death rate) is:

$$\frac{\text{The } \textit{total number} \text{ of deaths during a year in the population at risk}}{\text{The population at risk (usually taken as the population at midyear)}}$$

The cause-specific annual mortality rate is:

$$\frac{\text{The number of deaths occurring due to a particular cause during the year in the population at risk}}{\text{The population at risk (all those alive at midyear)}}$$

The age-specific annual mortality rate is:

$$\frac{\text{The number of deaths occurring in the given age group during the year in the population at risk}}{\text{The population at risk (all those alive at midyear)}}$$

A reason for taking the population at midyear as the denominator is that a population may grow or shrink during the year in question and the midyear population is an estimate of the average number during the year. One can, however, speak of death rates over a 5-year period rather than a 1-year period, and one can define the population at risk as all those alive at the beginning of the period.

4.3 Age-Adjusted Rates

When comparing death rates between two populations, the age composition of the populations must be taken into account. Since older people have a higher number of deaths per 1000 people, if a population is heavily weighted by older people, the crude mortality rate would be higher than in a younger population, and a comparison between the two groups might just reflect the age discrepancy rather than an intrinsic difference in mortality experience. One way to deal with this problem is to compare age-specific death rates, death rates specific to a particular age group. Another way

that is useful when an overall summary figure is required is to use *age-adjusted* rates. These are rates adjusted to what they *would be* if the two populations being compared had the same age distributions as some arbitrarily selected standard population.

For example, the table below shows the crude and age-adjusted mortality rates for the United States at five time periods 15.7. The adjustment is made to the age distribution of the population in 1940 as well as the age distribution of the population in 2000. Thus, we see that in 1991, the age-adjusted rate was 5.1/1000 when adjusted to 1940 standard, but the crude mortality rate was 8.6/1000. This means that if in 1991 the age distribution of the population were the same as it was in 1940, then the death rate would have been only 5.1/1000 people. The crude and age-adjusted rates for 1940 are the same because the 1940 population serves as the "standard" population whose age distribution is used as the basis for adjustment.

When adjusted to the year 2000 standard, the age-adjusted rate was 9.3. If in 1991 the age distribution were the same as in 2000, then the death rate would have been 9.3/1000 people. So, age-adjusted rates depend on the standard population being used for the adjustment. Note that the age-adjusted rate based on the population in year 2000 is higher than the age-adjusted rate based on the population in 1940; this is because the population is older in year 2000.

Year	Crude mortality rate per 1000 people	Age-adjusted rate (to population in 1940)	Age-adjusted rate (to population in 2000)
1940	10.8	10.8	17.9
1960	9.5	7.6	13.4
1980	8.8	5.9	10.4
1991	8.6	5.1	9.3
2001	8.5	Not computed after 1998	8.6

Although both crude and age-adjusted rates have decreased from 1940, the decrease in the age-adjusted rate is much greater. The percent change in crude mortality between 1940 and 1991 was (10.8–8.6)/10.8 = 20.4%, whereas the percent change in the age-adjusted rate was (10.8–5.1)/10.8 = 0.528 or 52.8%.

The reason for this is that the population is growing older. For instance, the proportion of persons 65 years and over doubled between 1920 and 1960, rising from 4.8% of the population in 1920 to 9.6% in 1969. After 1998, the National Center for Health Statistics used the population in 2000 as the standard population against which adjustments were made. The crude rate and the age-adjusted death rate in the year 2001 are similar, and that is because the age distribution in 2001 is similar to the age distribution in 2000 so age adjustment doesn't really change the mortality rate much.

The age-adjusted rates are fictitious numbers—they do not tell you how many people actually died per 1000, but how many *would have* died if the age compositions were the same in the two populations. However, they are appropriate for comparison purposes. Methods to perform age adjustment are described in Appendix 5.

4.4 Incidence and Prevalence

Prevalence and incidence are two measures of morbidity (illness).

Prevalence of a disease is defined as:

$$\frac{\text{The number of persons with a disease}}{\text{The total number of persons in population at risk at a particular point in time}}$$

(This is also known as *point prevalence*, but more generally referred to just as "prevalence.") For example, the prevalence of hypertension in 1973 among black males, ages 30–69, defined as a diastolic blood pressure (DBP) of 95 mmHg or more at a blood pressure screening program conducted by the Hypertension Detection and Follow-Up Program (HDFP),[16] was calculated to be

$$\frac{4,268\ \textit{with DBP} > 95\ \textit{mmHg}}{15,190\ \textit{black men aged}\ 30-69\ \textit{screened}} \times 100 = 28.1\,\text{per}\,100$$

Several points are to be noted about this definition:

(1) The risk group (denominator) is clearly defined as black men, ages 30–69.
(2) The point in time is specified as time of screening.
(3) The definition of the disease is clearly specified as a diastolic blood pressure of 95 mmHg or greater. (This may include people who are treated for the disease but whose pressure is still high and those who are untreated.)
(4) The numerator is the subset of individuals in the reference group (denominator) who satisfy the definition of the disease.

Prevalence can also be age-adjusted to a standard population, meaning that the prevalence estimates are adjusted to what they would be if the population of interest had the same age distribution as some standard population (see Appendix 5 and Section 4.3). The prevalence of hypertension estimated by the Hispanic Community Health Study/Study of Latinos (HCHS/SOL),[17] in 2007–2011, was found to be 25.5% when age adjusted to the year 2000 standard. Hypertension was defined as systolic blood pressure of 140 mmHg or above or diastolic blood pressure of 90 mmHg or above or on medications for high blood pressure.

Incidence is defined as:

$$\frac{\text{The number of new cases of a disease per unit of time}}{\text{The total number at risk in beginning of this time period}}$$

For example, studies have found that the 10-year incidence of a major coronary event (such as heart attack) among white men, ages 30–59, with diastolic blood pressure 105 mmHg or above at the time they were first seen, was found to be 183 per 1000.[18] This means that among 1000 white men, ages 30–59, who had

diastolic blood pressure above 105 mmHg at the beginning of the 10-year period of observation, 183 of them had a major coronary event (heart attack or sudden death) during the next 10 years. Among white men with diastolic blood pressure of <75 mmHg, the 10-year incidence of a coronary event was found to be 55/1000. Thus, comparison of these two incidence rates, 183/1000 for those with high blood pressure versus 55/1000 for those with low blood pressure, may lead to the inference that elevated blood pressure is a risk factor for coronary disease.

Often, one may hear the word "incidence" used when what is really meant is prevalence. You should beware of such incorrect usage. For example, you might hear or even read in a medical journal that the incidence of diabetes in 1973 was 42.6 per 1000 individuals, ages 45–64, when what is really meant is that the prevalence was 42.6/1000. The thing to remember is that prevalence refers to the *existence of a disease* at a specific period in time, whereas incidence refers to *new cases* of a disease developing within a specified period of time.

Note that *mortality rate is incidence*, whereas *morbidity may be expressed as incidence or prevalence.* In a chronic disease, the prevalence is greater than the incidence because prevalence includes both new cases and existing cases that may have first occurred a long time ago, but the afflicted patients continued to live with the condition. For a disease that is either rapidly fatal or quickly cured, incidence and prevalence may be similar. Prevalence can be established by doing a survey or a screening of a target population and counting the cases of disease existing at the time of the survey. This is a cross-sectional study. Incidence figures are harder to obtain than prevalence figures since to ascertain incidence, one must identify a group of people free of the disease in question (i.e., a cohort), observe them over a period of time, and determine how many develop the disease over that time period. The implementation of such a process is difficult and costly.

4.5 Standardized Mortality Ratio

The *standardized mortality ratio* (SMR) is the ratio of the number of deaths observed to the number of deaths expected. The number expected for a particular age group, for instance, is often obtained from population statistics.

$$SMR = \frac{observed\ \ deaths}{expected\ \ deaths}$$

4.6 Person-Years of Observation

Occasionally, one sees a rate presented as some number of events *per person-years of observation*, rather than per number of individuals observed during a specified period of time. Per person-years (or months) is useful as a unit of measurement when

people are observed for different lengths of time. Suppose you are observing cohorts of people free of heart disease to determine whether the incidence of heart disease is greater for smokers than for those who quit. Quitters need to be defined, for example, as those who quit more than 5 years prior to the start of observation. One could define quitters differently and get different results, so it is important to specify the definition. Other considerations include controlling for the length of time smoked, which would be a function of age at the start of smoking and age at the start of the observation period, the number of cigarettes smoked, and so forth. But for simplicity, we will assume everyone among the smokers has smoked an equal amount and everyone among the quitters has smoked an equal amount prior to quitting.

We can express the incidence rate of heart disease per some unit of time, say 10 years, as the number who developed the disease during that time, divided by the number of people we observed (number at risk). However, suppose we didn't observe everyone for the same length of time. This could occur because some people moved or died of other causes or were enrolled in the study at different times or for other reasons. In such a case, we could use as our denominator the number of *person-years of observation*.

For example, if individual 1 was enrolled at time 0 and was observed for 4 years and then lost to follow-up, he/she would have contributed 4 person-years of observation. Ten such individuals would contribute 40 person-years of observation. Another individual observed for 8 years would have contributed 8 person-years of observation, and 10 such individuals would contribute 80 person-years of observation for a total of 120 person-years. If six cases of heart disease developed among those observed, the rate would be 6 per 120 person-years, rather than 6/10 individuals observed. Note that if the denominator is given as person-years, you don't know if it pertains to 120 people each observed for 1 year, or 12 people each observed for 10 years or some combination. Another problem with this method of expressing rates is that it reflects the average experience over the time span, but it may be that the rate of heart disease is the same for smokers as for quitters within the first 3 years and the rates begin to separate after that. In any case, various statistical methods are available for use with person-year analysis.

4.7 Incidence Rate Ratio (IRR)

Incidence Rate Ratio (IRR) compares two groups on their incidence rates of a disease or condition (calculated in person-years).

$$\text{IRR} = \frac{(\text{IRRgroup}_\text{A})}{(\text{IRRgroup}_\text{B})}$$

To calculate the confidence interval for an IRR, you first need to get the standard error (SE) of the log of IRR.

$$\text{SE } \ln(\text{IRR}) = \sqrt{\frac{1}{n \text{ events in grp } A} + \frac{1}{n \text{ events in grp } B}}$$

The 95% confidence interval bounds are $\ln(\text{IRR}) \pm 1.96\ (\text{SE} \times \ln(\text{IRR}))$. But then you have to get out of the log scale and go back to the original numbers by taking the exponents, so

The upper confidence limit is:

$$e^{(\ln(\text{IRR}) + 1.96\ \text{SE}(\ln(\text{IRR})))}$$

The lower 95% confidence limit is:

$$e^{(\ln(\text{IRR}) - 1.96\ \text{SE}(\ln(\text{IRR})))}$$

4.8 Dependent and Independent Variables

In research studies, we want to quantify the relationship between one set of variables, which we may think of as predictors or determinants, and some outcome or criterion variable in which we are interested. This outcome variable, which it is our objective to explain, is the dependent variable.

A *dependent variable* is a factor whose value depends on the level of another factor, which is termed an *independent variable*. In the example of cigarette smoking and lung cancer mortality, the duration and number of cigarettes smoked are independent variables upon which the lung cancer mortality depends (thus, lung cancer mortality is the dependent variable).

4.9 Types of Studies

In Section 1.4, we described different kinds of study designs, in the context of our discussion of the scientific method and of how we know what we know. These were observational studies, which may be cross-sectional, case–control, or prospective and experimental studies, which are clinical trials. In the following sections, we will consider the types of inferences that can be derived from data obtained from these different designs.

The objective is to assess the relationship between some factor of interest (the independent variable), which we will sometimes call exposure, and an outcome variable (the dependent variable).

The observational studies are distinguished by the *point in time when measurements are made on the dependent and independent variables*, as illustrated below. In cross-sectional studies, both the dependent and independent (outcome and exposure) variables are measured at the same time, in the present. In case–control studies, the

outcome is measured now and exposure is estimated from the past. In prospective studies, exposure (the independent variable) is measured now and the outcome is measured in the future. In the next section, we will discuss the different inferences to be made from cross-sectional versus prospective studies.

	Time of measurement		
	Past	Present	Future
Cross-sectional		Exposure outcome	
Case–control	Exposure	Outcome	
Prospective		Exposure	Outcome

4.10 Cross-Sectional Versus Longitudinal Looks at Data

Prospective studies are sometimes also known as longitudinal studies, since people are followed longitudinally, over time. Examination of longitudinal data may lead to quite different inferences than those to be obtained from cross-sectional looks at data. For example, consider age and blood pressure.

Cross-sectional studies have repeatedly shown that the average systolic blood pressure is higher in each successive 10-year age group, while diastolic pressure increases for age groups up to age 50 and then reaches a plateau. One cannot, from these types of studies, say that blood pressure rises with age because the pressures measured for 30-year-old men, for example, were not obtained on the same individuals 10 years later when they were 40, but were obtained for a different set of 40-year-olds. To determine the effect of age on blood pressure, we would need to take a longitudinal or prospective look at the same individuals as they get older. One interpretation of the curve observed for diastolic blood pressure, for instance, might be that individuals over 60 who had very high diastolic pressures died off, leaving only those individuals with lower pressure alive long enough to be included in the sample of those having their blood pressure measured in the cross-sectional look.

The diagrams in Figure 4.2 illustrate the possible impact of a "cohort effect," a cross-sectional view, and a longitudinal view of the same data. (Letters indicate groups of individuals examined in a particular year.)

If you take the blood pressure of all groups in 2010 and compare group F to group D, you will have a cross-sectional comparison of 30-year-olds with 40-year-olds at a given point in time. If you compare group F in 2010 with group F (same individuals) in 2020, you will have a longitudinal comparison. If you compare group F in 2010 with group H in 2020, you will have a comparison of blood pressures of 30-year-olds at different points in time (a horizontal look).

These comparisons can lead to quite different conclusions, as is shown by the schematic examples in Figure 4.3 using fictitious numbers to represent average diastolic blood pressure.

In example (1), measurements in 2010 indicate that average diastolic blood pressure for 30-year-olds (group F) was 90 mmHg and for 40-year-olds (group D)

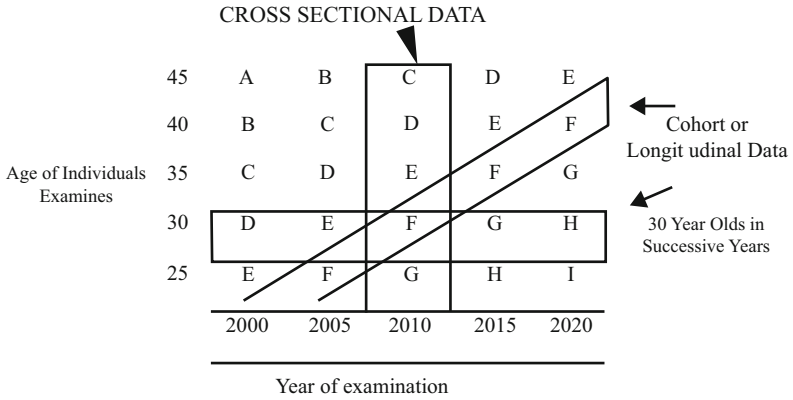


Figure 4.2 Cross-sectional and Longitudinal Effects

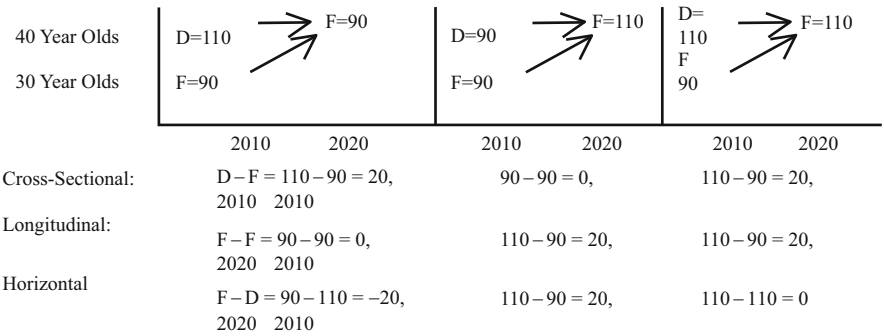


Figure 4.3 Cross-sectional versus longitudinal comparisons

it was 110 mmHg. Looking at group F 10 years later, when they were 40-year-olds, indicates their mean diastolic blood pressure was 90 mmHg. The following calculations result:

Cross-sectional look	$D_{2010} - F_{2010} = 110 - 90 = 20$
Conclusion	40-year-olds have higher blood pressure than 30-year-olds (by 20 mmHg)
Longitudinal look	$F_{2020} - F_{2010} = 90 - 90 = 0$
Conclusion	Blood pressure does not rise with age
Horizontal look (cohort comparisons)	$F_{2020} - F_{2010} = 90 - 110 = -20$
Conclusion	40-year-olds in 2020 have lower blood pressure than 40-year-olds did in 2010

A Possible Interpretation Blood pressure does not rise with age, but different environmental forces were operating for the F cohort than for the D cohort.

In example (2), we have

Cross-sectional look	D – F = 90–90 = 0 mmHg 2010 2010
Conclusion	From cross-sectional data, we conclude that blood pressure is not higher with older age
Longitudinal look	F – F = 110–90 = 20 2020 2010
Conclusion	From longitudinal data, we conclude that blood pressure goes up with age
Horizontal look	F – D = 110–90 = 20 2020 2010
Conclusion	40-year-olds in 2020 have higher blood pressure than 40-year-olds in 2010

A Possible Interpretation Blood pressure does rise with age and different environmental factors operated on the F cohort than on the D cohort.

In example (3), we have

Cross-sectional look	D – F = 110–90 = 20 2010 2010
Conclusion	Cross-sectionally, there was an increase in blood pressure for 40-year-olds over that for 30-year-olds
Longitudinal look	F – F = 110–90 = 20 2020 2010
Conclusion	Longitudinally, it is seen that blood pressure increases with increasing age
Horizontal look	F – D = 110–110 = 0 2020 2010
Conclusion	There was no change in blood pressure among 40-year-olds over the 10-year period

A Possible Interpretation Blood pressure does go up with age (supported by both longitudinal and cross-sectional data), and environmental factors affect both cohorts similarly.

4.11 Measures of Relative Risk: Inferences from Prospective Studies (the Framingham Study)

In epidemiologic studies, we are often interested in knowing how much more likely an individual is to develop a disease if he or she is exposed to a particular factor than the individual who is not so exposed. A simple measure of such likelihood is called *relative risk (RR)*. It is the ratio of two incidence estimates: *the rate of development of the disease for people with the exposure factor, divided by the rate of development*

of the disease for people without the exposure factor. Suppose we wish to determine the effect of high blood pressure (hypertension) on the development of cardiovascular disease (CVD). To obtain the relative risk, we need to calculate the incidence rates. We can use the data from a classic prospective study, the Framingham Heart Study.[19]

This was a pioneering prospective epidemiologic study of a population sample in the small town of Framingham, Massachusetts. Beginning in 1948, a *cohort* of people was selected to be followed up biennially. The term *cohort* refers to a group of individuals followed longitudinally over a period of time. A birth cohort, for example, would be the population of individuals born in a given year. The Framingham cohort was a sample of people chosen at the beginning of the study period and included men and women aged 30–62 years at the start of the study. These individuals were observed over a 20-year period, and morbidity and mortality associated with cardiovascular disease were determined. A standardized hospital record and death certificate were obtained, clinic examination was repeated at 2-year intervals, and the major concern of the Framingham study has been to evaluate the relationship of characteristics determined in *well* persons to the subsequent development of disease.

Through this study, "risk factors" for cardiovascular disease were identified. The *risk factors* are antecedent physiological characteristics or dietary and living habits, whose presence increases the individual's probability of developing cardiovascular disease at some future time. Among the most important predictive factors identified in the Framingham study were *elevated blood pressure*, *elevated serum cholesterol*, *and cigarette smoking*. Elevated blood glucose and abnormal resting electrocardiogram findings are also predictive of future cardiovascular disease.

Relative risk can be determined by the following calculation:

$$\frac{\begin{array}{l}\text{Incidence of cardiovascular disease (new cases)}\\ \text{over a specified period of time among people free}\\ \text{of CVD at beginning of the study period who have}\\ \text{the risk factor in question (e.g., high blood pressure)}\end{array}}{\begin{array}{l}\text{Incidence of CVD in the given time period among}\\ \text{peoplefree of CVD initially, who do not have the risk}\\ \text{factor in question (normal blood pressure)}\end{array}}$$

From the Framingham data, we calculate for men in the study the

$$\text{RR of CVD within 18 years after first exam} = \frac{353.2/10,000 \text{ persons at risk with definite hypertension}}{123.9/10,000 \text{ persons at risk with no hypertension}}$$

$$\frac{353.2}{123.9} = 2.85$$

This means that a man with definite hypertension is 2.85 times more likely to develop CVD in an 18-year period than a man who does not have hypertension. For women, the relative risk is

$$\frac{187.9}{57.3} = 3.28$$

This means that hypertension carries a somewhat greater relative risk for women. But note that the *absolute* risk for persons with definite hypertension (i.e., the incidence of CVD) is greater for men than for women, being 353.2 per 10,000 men versus 187.9 per 10,000 women.

The incidence estimates given above have been age adjusted. Age adjustment is discussed in Section 4.3. Often, one may want to adjust for other variables such as smoking status, diabetes, cholesterol levels, and other factors that may also be related to outcome. This may be accomplished by multiple logistic regression analysis and by Cox proportional hazards analysis, which are described in Sections 4.18 and 4.20, respectively, but first we will describe how relative risk can be calculated from prospective studies or estimated from case–control studies.

4.12 Calculation of Relative Risk from Prospective Studies

Relative risk can be determined directly from prospective studies by constructing a 2 × 2 table as follows[20]:

		DISEASE (developing in the specified period)		
		Yes	No	
RISK FACTOR (determined at beginning of study period)	PRESENT (high blood pressure)	$a = 90$	$b = 403$	$a + b = 493$ (persons with factor)
	ABSENT (normal blood pressure)	$c = 70$	$d = 1201$	$c + d = 1271$ (persons without factor)

Relative risk is

$$\frac{\textit{incidence of disease among those with high BP}}{\textit{incidence disease among those with normal BP}} = \frac{a/(a+b)}{c/(c+d)} = \frac{90/493}{70/1271} = 3.31$$

Relative risk, or hazard ratio, can be calculated from Cox proportional hazards regression models (which allow for adjustment for other variables) as described in Section 4.20.

4.13 Odds Ratio: Estimate of Relative Risk from Case–Control Studies

A case–control study is one in which the investigator seeks to establish an association between the presence of a characteristic (a risk factor) and the occurrence of a disease *by starting out with a sample of diseased persons and a control group of nondiseased persons and by noting the presence or absence of the characteristic in each of these two groups.* It can be illustrated by constructing a 2 × 2 table as follows:

		DISEASE	
		PRESENT	ABSENT
RISK FACTOR	PRESENT	a	b
	ABSENT	c	d
		$a + c$ (number of persons *with* disease)	$b + d$ (number of persons *without* disease)

The objective is to determine if the proportion of persons with the disease who have the factor is greater than the proportion of persons without the disease who have the factor. In other words, it is desired to know whether $a/(a + c)$ is greater than $b/(b + d)$.

Case–control studies are often referred to as *retrospective studies* because the investigator must gather data on the *independent* variables *retrospectively.* The dependent variable—the presence of disease—is obtained at time of sampling, in contrast to prospective studies where the independent variables are measured at time of sampling and the dependent variable is measured at some future time (i.e., when the disease develops). The real distinction between case–control or retrospective studies and prospective studies has to do with selecting individuals for the study—those *with and without the disease* in case–control/retrospective studies and those *with and without the factor* of interest in prospective studies.

Since in prospective studies *we sample the people with the characteristic of interest and the people without the characteristic*, we can obtain the relative risk *directly* by calculating the incidence rates of disease in these two groups. In case–control studies, however, *we sample patients with and without the disease* and note the presence or absence of the characteristic of interest in these two groups; we do

not have a direct measure of *incidence* of disease. Nevertheless, making certain assumptions, we can make some approximations to what the relative risk would be if we could measure incidence rates through a prospective study. These approximations hold best for diseases of *low incidence.*

To estimate relative risk from case–control studies, note that

$$\frac{a/(a+b)}{c/(c+d)} = \frac{a\ \ (c+d)}{c\ \ (a+b)}$$

Now assume that since the disease is not all that common, c is negligible in relation to d (in other words, among people without the risk factor, there aren't all that many who will get the disease, relative to the number of people who will not get it). Assume also that, *in the population*, a is negligible relative to b, since even among people with the risk factor, not all that many will get the disease in comparison to the number who won't get it. Then, the terms in the parentheses become d in the numerator and b in the denominator so that

$$\frac{a\ \ (c+d)}{c\ \ (a+b)} \textit{reduces to } OR = \frac{ad}{bc}$$

This is known as the *odds ratio* (OR) and is a good estimate of relative risk when the disease is rare.

An example of how the odds ratio is calculated is shown below. In a case–control study of lung cancer, the table below was obtained.[21] Note that we are not sampling smokers and nonsmokers here. Rather, we are sampling those with and without the disease. So, although in the *population at large a* is small relative to b, in this sample it is not so.

	Patients with lung cancer		Matched controls with other diseases	
Smokers of 15–24 cigarettes daily	475	a	431	b
Nonsmokers	7	c	61	d
	(persons with disease)		(persons without disease)	

The odds ratio, as an estimate of the relative risk of developing lung cancer for people who smoke 15–24 cigarettes a day, compared with nonsmokers is

$$\textit{Odds ratio} = \frac{475 \times 61}{431 \times 7} = 9.60 = \textit{estimate of relative risk}$$

This means that smokers of 15–24 cigarettes daily are 9.6 times more likely to get lung cancer than are nonsmokers.

One more thing about the odds ratio is that it is the ratio of odds of lung cancer for those who smoke 15–24 cigarettes a day, relative to odds of lung cancer for those who don't smoke. In the example above, we get

$$\begin{gathered} for\ \ smokers:\ \ odds\ \ of\ \ lung\ \ cancer\ \ are \frac{475}{431} \\ for\ \ nonsmokers: odds\ \ of\ \ lung\ \ cancer\ \ are \frac{7}{61} \\ ratio\ \ of\ \ odds = \frac{475/431}{7/61} \end{gathered}$$

So the point is that the odds ratio is the odds ratio, whether the disease is rare or not. It is always the ratio of odds of disease for those with the exposure versus the odds of disease for those without the exposure. But when the disease is rare, it is also a good estimate of the relative risk.

We can also put confidence limits on the odds ratio. We calculated the odds ration as 9.60. The 95% confidence limits for an odds ratio (OR) are

$$OR \times e^{\left[\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]}$$

Upper 95% confidence limit =

$$OR \times e^{\left[1.96\sqrt{\frac{1}{475}+\frac{1}{431}+\frac{1}{7}+\frac{1}{61}}\right]}$$

$$OR \times e^{(1.96 \times .405)} = 9.6 \times e^{.794} = 9.6 \times 2.21 = 21.2$$

Lower 95% confidence limit =

$$\begin{gathered} OR \times e^{(-1.96 \times .405)} = 9.6 \times e^{-.794} = 9.6 \times .45 = 4.3 \\ Note:\ \ e^{-.794} = \frac{1}{e^{.794}} = .45 \end{gathered}$$

Thus, the confidence *interval* is 4.3–21.2.

We often express this as (OR; 95% confidence limits), which is our example:

$$(\text{OR} = 9.60;\ \ 95\%\text{CI}:\ \ 4.3,\ \ 21.2)$$

Odds ratios can be calculated from logistic regression (which allow for adjustment for other variables) as described in Section 4.18.

4.14 Attributable Risk

Attributable risk (AR) is:

$$\text{The risk in exposed—risk in unexposed individuals.}$$

Population attributable risk (PAR) is:

$$\text{AR} \times \text{risk factor prevalence}$$

While relative risk pertains to the risk of a disease in exposed persons *relative to the risk* in the unexposed, the attributable risk pertains to *the difference in absolute risk* of the exposed compared to the unexposed persons. It may tell us how much *excess risk* there is due to the exposure in the exposed. In the example in Section 4.12, the 10-year risk among those with high blood pressure was 90/493 = 0.183 (or 183 per 1000 people with high blood pressure), while in those with normal blood pressure, it was 70/1271 = 0.055 (or 55 per 1000 with normal pressure).

Thus, the attributable risk in those exposed (i.e., with high blood pressure) is 0.183–0.055 = 0.128 (128 per 1000). In other words, heart disease events in 128 of the 183 people per 1000 with high blood pressure can be attributed to the high blood pressure. We can also express this excess as a percentage of the risk in the exposed that is attributable to the exposure

$$\frac{128/1000}{183/1000} = 128 = .70 \ \ or \ \ 70\%$$

But, we must be very careful about such attribution—it is only valid when we can assume the exposure causes the disease (after taking into account confounding and other sources of bias).

Population attributable risk (PAR) is a useful measure when we want to see how we could reduce morbidity or mortality by eliminating a risk factor. It depends on the prevalence of the risk factor in the population as noted above. Here is an example from the Women's Health Initiative (described in more detail in Chapter 6). It was found in a clinical trial that postmenopausal women who were taking estrogen plus progestin had an annualized rate of coronary heart disease of 39 per 10,000 compared to a rate of 33 per 10,000 for women taking placebo.[22]

Thus,

$$AR = \frac{39 - 33}{10,000} = .0006$$

or six excess coronary heart disease events per 10,000 women taking this preparation, per year.

Since approximately 6,000,000 women were taking that hormone preparation at the time (exposed), then 0.0007 × 6,000,000 = 3600 coronary heart disease events per year could be *attributed* to taking estrogen plus progestin.

The prevalence of use of estrogen plus progestin estimated from the same study when it was first begun was about 18%. If we use this estimate,

$$\text{PAR} = \text{AR} \times \text{prevalence of risk factor} = .0006 \times .18 = .000108;$$

Thus, if use of estrogen plus progestin were eliminated, there would be 10.8 per 100,000 fewer postmenopausal women who had heart disease events.

4.15 Response Bias

There are many different types of bias that might lead you to either underestimate or overestimate the size of a relative risk of odds ratio, and it is important to try to anticipate potential sources of bias and avoid them. The illustration on the next page shows the impact of one kind of possible bias: *ascertainment or response bias.*

Assume that you have the following situation. Of 100 people exposed to a risk factor, 20% develop the disease, and of a 100 people unexposed, 16% develop the disease, yielding a relative risk of 1.25, as shown in the illustration.

Now imagine that only 60% of the exposed respond to follow-up or are ascertained as having or not having the disease, *a 60% response rate among the exposed*. Assume further that all of the ones who don't respond happen to be among the ones who *don't* develop disease. The relative risk would be calculated as 2.06.

Now imagine that only 60% of the nonexposed reply, *a 60% response rate among the nonexposed*, and all of the nonexposed who don't respond happen to be among the ones who don't have the disease. Now, the relative risk estimate is 0.75.

To summarize, you can get conflicting estimates of the relative risk if you have differential response rates. Therefore, it is very important to get as complete a response or ascertainment as possible. The tables showing these calculations follow.

FULL RESPONSE RATE

		DISEASE +	DISEASE –	
EXPOSURE	+	R = 100% 20	100% 80	100
	–	100% 16	100% 84	100
		36	164	200

$$RR = \frac{20 / 100}{16 / 100} = \frac{.20}{.16} = 1.25$$

DIFFERENTIAL RESPONSE RATE

EXPOSURE	DISEASE +	DISEASE –	
+	R = 100% 20	50% 40	60 (response rate = 60%)
–	100% 16	100% 84	100
	36	124	160

$$RR = \frac{20/60}{16/100} = \frac{.33}{.16} = 2.06$$

DIFFERENTIAL RESPONSE RATE

EXPOSURE	DISEASE +	DISEASE –	
+	R = 100% 20	100% 40	100
–	100% 16	52% 84	60 (response rate = 60%)
	36	124	160

$$RR = \frac{20/100}{16/60} = \frac{.20}{.266} = .75$$

4.16 Confounding Variables

A *confounding variable* is one that is closely associated with both the independent variable and the outcome of interest in those unexposed. For example, a confounding variable in studies of coffee and heart disease may be smoking. Since some coffee drinkers are also smokers, if a study found a relationship between coffee drinking (the independent variable) and development of heart disease (the dependent

variable), it could really mean that it is the smoking that causes heart disease, rather than the coffee. In this example, smoking is the confounding variable.

If *both* the confounding variable and the independent variable of interest are closely associated with the dependent variable, then the observed relationship between the independent and dependent variables may really be a reflection of the *true* relationship between the confounding variable and the dependent variable. An intuitive way to look at this is to imagine that if a confounding variable were perfectly associated with an independent variable, it could be substituted for it. It is important to account or adjust for confounding variables *in the design and statistical analysis of studies* in order to avoid wrong inferences.

There are several approaches to dealing with potential confounders. One approach is to deal with it in the study design by matching, for example, as described in Section 4.17; another way of controlling for confounding variables is in the data analysis phase, by using multivariate analysis, as described in the sections below. An excellent discussion is found in *Modern Epidemiology* by Rothman, Lash, and Greenland.

4.17 Matching

One approach to dealing with potential confounders is to match subjects in the two groups on the confounding variable. In the example discussed above concerning studies of coffee and heart disease, we might match subjects on their smoking history, since smoking may be a confounder of the relationship between coffee and heart disease. Whenever we enrolled a coffee drinker into the study, we would determine if that person was a smoker. If the patient was a smoker, the next patient who would be enrolled who was not a coffee drinker (i.e., a member of the comparison group) would also have to be a smoker. For each coffee-drinking nonsmoker, a non-coffee-drinking nonsmoker would be enrolled. In this way, we would have the same number of smokers in the two groups. This is known as *one-to-one matching*. There are other ways to match, and these are discussed more fully in the books by Anderson, Elwood, and Rothman as noted in the "Suggested Readings" section.

In case–control studies, finding an appropriate comparison group may be difficult. For example, suppose an investigator is studying the effect of coffee on pancreatic cancer. The investigator chooses as control patients in the hospital at the same time and in the same ward as the cases but with a diagnosis other than cancer. It is possible that patients hospitalized for gastrointestinal problems other than cancer might have voluntarily given up coffee drinking because it bothered their stomachs. In such a situation, the coffee-drinking habits of the two groups might be similar, and the investigator might not find a greater association of coffee drinking with cases than with controls. A more appropriate group might be patients in a different ward, say an orthopedic ward. But here one would have to be careful to match on age, since orthopedic patients may be younger than the other cases if the

hospital happens to be in a ski area, for example, where reckless skiing leads to broken legs, or they may be substantially older than the other cases if there are many patients with hip replacements due to falls in the elderly or osteoarthritis.

It needs to be pointed out that the factor that is matched cannot be evaluated in terms of its relationship to outcome. Thus, if we are comparing two groups of women for the effect of vitamin A intake on cervical cancer and we do a case–control study in which we enroll cases of cervical cancer and controls matched on age, we will not be able to say from this study whether age is related to cervical cancer. This is because we have ensured that the age distributions are the same in both the case and control groups by matching on age, so obviously we will not be able to find differences in age between the groups.

Some statisticians believe that matching is often done unnecessarily and that if you have a large enough study, simple randomization or stratified randomization is adequate to ensure a balanced distribution of confounding factors. Furthermore, multivariate analysis methods, such as logistic regression or proportional hazards models, provide another, usually better, way to control for confounders. A good discussion of matching can be found in the book *Methods in Observational Epidemiology*, by Kelsey and colleagues.

4.18 Multiple Logistic Regression

Multiple logistic regression analysis is used to calculate the probability of an event happening as a function of several independent variables. It is useful in controlling for confounders when examining the relationship between an independent variable and the occurrence of outcome (e.g., such as heart attack) within a specified period of time. The equation takes the form of

$$P(event) = \frac{1}{1 + e^{-k}}$$

where $k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots.. + \beta_m x_m$(note that $e^{-k} = 1/e^k$)

Each x_i is a particular independent variable and the corresponding coefficients, C's, are calculated from the data obtained in the study. For example, let us take the Framingham data for the probability of a man developing cardiovascular disease within 8 years. Cardiovascular disease (CVD) was defined as coronary heart disease, brain infarction, intermittent claudication, or congestive heart failure.

$$P(\text{CVD}) = \frac{1}{1 + e^{\left[-19.77 + .37(age) - .002(age)^2 + .026(chl) + .016(SBP) + .558(SM) + 1.053(LVH) + .602(GI) - .00036(chl \times age)\right]}}$$

where:

chl = serum cholesterol
SBP = systolic blood pressure
SM = 1 if yes for smoking, 0 if no
LVH = left ventricular hypertrophy, 1 if yes, 0 if no
Gl = glucose intolerance, 1 if yes, 0 if no

For example, suppose we consider a 50-year-old male whose cholesterol is 200 and systolic blood pressure is 160, who smokes and who has no LVH and no glucose intolerance. When we multiply the coefficients by this individual's values on the independent variables and do the necessary calculations, we come up with a probability of 0.17. This means that this individual has 17 chances in a 100 of developing some form of cardiovascular disease during a period of 8 years.

The coefficients from a multiple logistic regression analysis can be used to calculate the odds ratio for one factor while controlling for all the other factors. The way to do this is to take the natural log e raised to the coefficient for the variable of interest, if the variable is a dichotomous one (i.e., coded as 1 or 0). For example, the odds of cardiovascular disease for smokers relative to nonsmokers among males, while controlling for age, cholesterol, systolic blood pressure, left ventricular hypertrophy, and glucose intolerance, is $e^{.558} = 1.75$. This means that a person who smokes has 1.75 times higher odds of getting CVD (within 8 years) than the one who doesn't smoke if these two individuals are equal with respect to the other variables in the equation. This is equivalent to saying that the smoker's risk is 75% higher than the nonsmokers.

If we want to compare the odds of someone with a systolic blood pressure of 200 versus someone with systolic blood pressure of 120, all other factors being equal, we calculate it as follows:

$$OR = e^{\beta(200-120)} = e^{.016(80)} = e^{1.28} = 3.6$$

The man with systolic blood pressure of 200 mmHg is 3.6 times more likely to develop disease during the 8-year period than the one with pressure of 120. Note that we are talking about odds ratios and we think of them as approximations of relative risk.

Thus, there are two quantities we can obtain from a logistic regression: (1) the probability of an event occurring within a specified period of time, for a person with a particular set of characteristics, and (2) the odds ratio of the event occurring in a person with one value of an independent variable compared to a person with a different value while controlling for all the other variables.

Logistic regression can be used in case–control studies. Raising e to the coefficient of the variable of interest gives us the odds ratio. The confidence intervals for the odds ratio are

$$\text{OR} = \text{e}^{[\beta - 1.96(s.e.)]} \text{ as the lower limit}$$
$$\text{OR} = \text{e}^{[\beta + 1.96(s.e.)]} \text{ as the upper limit.}$$

s.e. is the standard error of the beta coefficient and is an output of all statistical computer packages.

Multiple logistic regression is appropriate for cross-sectional and case–control studies when the dependent variable (outcome) is dichotomous (i.e., can be coded as 1 = event, 0 = no event) and when the question deals with the occurrence of the event of interest within a specified period of time and the people are all followed for that length of time. However, when follow-up time for people in the study differs, then survival analysis should be used, as described in Sections 4.19 and 4.20.

4.19 Survival Analysis: Life Table Methods

Survival analysis of data should be used when the follow-up times differ widely for different people or when they enter the study at different times. It can get rather complex, and this section is intended only to introduce the concepts. Suppose you want to compare the survival of patients treated by two different methods and suppose you have the data shown below.[23] We will analyze it by using the Kaplan–Meier survival curves.

Deaths at a given month in two groups

Status:	D L D D D		D D D D L
Group A:	4, 5+ 9, 11, 12	Group B:	2, 3, 4, 5, 6+

Status: (D = dead at that month; L = living at that month)
(The + means patient was lost to follow-up and last seen alive at that month)

In each group, four patients had died by 12 months, and one was seen alive some time during that year, so we don't know whether that patient was dead or alive at the end of the year. If we looked at the data in this way, we would have to say that the survival by 1 year was the same in both groups.

	Group	
At end of 12 months	**A**	**B**
Dead	4	4
Alive	1	1
Survival rate	20%	20%

However, a more appropriate way to analyze such data is through *survival curves*. The points for the curves are calculated as shown in the table below.

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7
Case#	Time in Mos.	Status	# pts. enter	Prop. Dead $q_i = \frac{\text{dead}}{\text{entered}}$	Prop. Surv. $P_1 = 1—q_i$	Cum. Surv. $P_1 = P_{i—1} \times Pi$
Group A						
1	4	Dead	5	1/5 = 0.2	0.80	1 × 0.8 = 0.8
2	5	Surv	4	0/4 = 0.0	1.00	0.8 × 1 = 0.8
3	9	Dead	3	1/3 = 0.33	0.67	0.8 × 0.67 = 0.53
4	11	Dead	2	1/2 = 0.5	0.50	0.53 × 0.5 = 27
5	12	Dead	1	1/1 = 1.0	0.00	0.27 × 0 = 0
Group B						
1	2	Dead	5	1/5 = 0.2	0.80	1 × 0.8 = 0.8
2	3	Dead	4	1/4 = 0.25	0.75	0.8 × 0.75 = 0.6
3	4	Dead	3	1/3 = 0.33	0.67	0.6 × 0.67 = 0.4
4	5	Dead	2	1/2 = 0.5	0.50	0.4 × 0.5 = 0.2
5	6	Surv	1	0/1 = 0.0	1.00	0.2 × 0.1 = 0.2

First of all, the patients are placed in order of the time of their death or the last time they were seen alive. Let us go through the third row for group A, as an example.

The third patient died at 9 months (columns 1 and 2). At the beginning of the 9th month, there were three patients at risk of dying (out of the total of five patients who entered the study). This is because one of the five patients had already died in the 4th month (case #1), and one was last seen alive at the 5th month (case #2) and so wasn't available to be observed. Out of these three patients at risk in the beginning of the 9th month, one died (case #3). So, the probability of dying in the 9th month is 1/3, and we call this q_i, where i in this case refers to the 9th month. Therefore, the proportion surviving in the 9th month is $p_i = 1—q_i = 0.1–0.33 = 0.67$.

The cumulative proportion surviving means the proportion surviving up through the 9th month. To survive through the 9th month, a patient had to have survived to the end of month 8 *and* have survived in month 9. Thus, it is equal to the cumulative probability of surviving *up to* the 9th month, which is 0.8, from column 7 row 2, *and* surviving in the 9th month, which is 0.67. We multiply these probabilities to get $0.8 \times 0.67 = 0.53$ as the probability of surviving through the 9th month. If we plot these points as in Figure 4.4, we note that the two survival curves look quite different and that group A did a lot better.

Survival analysis gets more complicated when we assume that patients who have been lost to follow-up in a given interval of time would have died at the same rate as those patient on whom we had information. Alternatively, we can make the calculations by assuming they all died within the interval in which they were lost to follow-up or they all survived during that interval.

Survival analysis can also be done while controlling for confounding variables, using the Cox proportional hazards model.

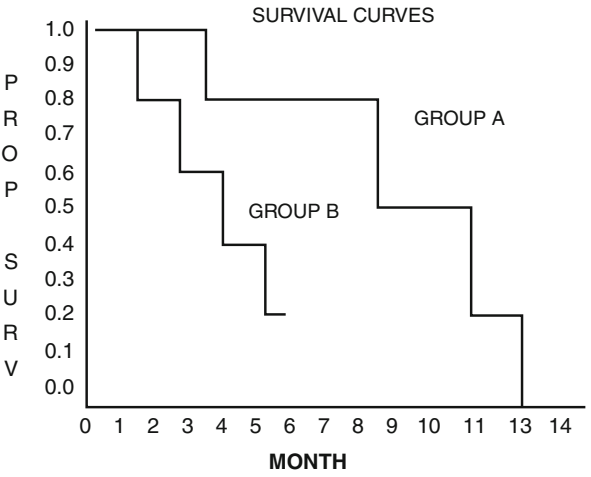


Figure 4.4 Survival curves

4.20 Cox Proportional Hazards Model

The Cox proportional hazards model is a form of multivariate survival analysis that can control for other factors. *The dependent variable is time to event (or survival time)*, which could be death, heart attack, or any other event of interest. This is in contrast to multiple logistic regression, where the dependent variable is a yes or no variable.

Cox proportional hazards model is appropriately used when there are different follow-up times because some people have withdrawn from the study or can't be contacted. People falling into one of those categories are considered to have "censored" observations. If the event of interest is, say, stroke, then people who died during the study from accidental causes would also be "censored" because we couldn't know whether they would have gone on to have a stroke or not, had they lived to the end of the study.

The coefficients from this analysis can be used to calculate an estimate of the relative risk of event, after controlling for the other covariates in the equation. This estimate from Cox proportional hazards models is more accurately called the "hazard ratio" or HR. If the event of interest is death, then it is the hazard at a point in time of dying in one group versus the hazard of dying in the other group. The ratio of these two hazards is the HR. The proportional hazards assumption means that we assume the ratio of these two hazards is the same over time.

An example of how to interpret results from such an analysis is given from the Systolic Hypertension in the Elderly Program (SHEP). This was a study of 4736 persons over age 60 with isolated systolic hypertension (i.e., people with high systolic blood pressure and normal diastolic blood pressure) to see if treatment with a low-dose diuretic and/or beta-blocker would reduce the rate of strokes compared with the rate in the control group treated with placebo.

A sample of a partial computer printout of a Cox regression analysis from the SHEP study is shown below. The event of interest is stroke in the placebo group.

Independent variable	Beta coefficient	s.e.	$e^{\text{Beta}} = \text{RR}$
Race	−0.1031	0.26070	0.90
Sex (male)	0.1707	0.19520	1.19
Age	0.0598	0.01405	1.06
History of diabetes	0.5322	0.23970	1.70
Smoking (baseline)	0.6214	0.23900	1.86

Let us look at the history of diabetes. The HR $= e^{.5322} = 1.70$, which is the natural logarithm e raised to the power specified by the beta coefficient; $e = 2.7183$. (Don't ask why.) This means that a person with untreated systolic hypertension who has a history of diabetes has 1.7 times the risk of having a stroke than a person with the same other characteristics but no diabetes. This can also be stated as a 70% greater risk. The 95% confidence limits for the hazard ratios are 1.06 and 2.72, meaning that we are 95% confident that the relative risk of stroke for those with a history of diabetes lies within the interval between 1.06 and 2.72. The formula for the 95% confidence interval for the hazard ratio is

$$\text{Limit 1} = e^{[\text{beta} - 1.96(\text{S.E.})]}$$
$$\text{Limit 2} = e^{[\text{beta} + 1.96(\text{S.E.})]}$$

If we are dealing with a continuous variable, like age, the HR is given per one unit or 1-year age increase. The hazard ratio per 5-year increase in age is

$$e^{5\ \text{beta}} = e^{5 \times .0598} = 1.35$$

There is a 34% increase in risk of future stroke per 5-year greater age at baseline, controlling for all the other variables in the model. To calculate confidence intervals for this example, you also need to multiply the s.e. by 5 (as well as multiplying the beta by 5), so the 95% confidence intervals of HR are [1.18, 1.55] for a 5-year increase in age.

The above example pertains to a linear relationship between age and risk of stroke. However, some relationships are not linear but rather have a J or U shape. Section 4.26 provides information on exploring a J- or U-shaped relationship between a variable and the outcome.

4.20.1 Difference Between Relative Risk and Hazards Ratio

It should be noted that the hazard ratio is not the same as a relative risk, though it is often thought of that way. A hazard rate is the instantaneous event rate at time t_e, given that it has not occurred before time t_e. We can write it as.

$$\text{Hazard} = \text{instantaneous event rate at } t_e \mid \text{no event at } t < t_e.$$

What we are interested in however is the hazard ratio. This is the ratio of the hazard in one group compared to the hazard in a comparison group, at a single point in time, or the ratio of two instantaneous hazards. Thus, if we say the hazard ratio of lung cancer for heavy smokers compared to never smokers is 19.9, we mean that heavy smokers are nearly 20 times as likely to get lung cancer as never smokers. If we are considering a continuous variable, like blood pressure, for example, we must specify per what unit we are talking about. For example, a hazard ratio of stroke for a 10 mm increase in daytime systolic blood pressure is 1.12; this means that the probability of stroke is 12% higher for each 10 mm increase in systolic blood pressure. We write this as:

$$\text{HR}_{10\text{mmHG}} = 1.12\ (95\%\text{CI} : 1.06 \text{ to } 1.18; p = 0.001).$$

The hazard ratio (HR) is the ratio of the instantaneous event rate (or disease rate) of those exposed to those not exposed at time t, provided they survived up to time t. The relative risk is the ratio of risk (probability) of an event in the exposed to the unexposed among those free of the event or disease at baseline. So, the populations used in these two estimates are different. In the case of HR, the population is those who survived to time t; in relative risk it's the population who were free of the disease at baseline.

4.20.2 Testing Cox Proportional Hazards Assumptions

Cox proportional hazards models assume that the effect of a covariate does not change over time, i.e., that the hazard ratio is constant over time. A common way to see if this assumption is met is by doing a visual inspection of *Schoenfeld residuals*. Here is how you would go about calculating Schoenfeld residuals (although in practice, you wouldn't calculate these since computer packages do it as part of Cox regressions, but this will give you an idea of what they mean):

(1) For each individual in your sample, you need to know whether the person had an event or not, i.e., whether $Y = 0$ or $Y = 1$, and if $Y = 1$, then the time when the event occurred t_j; the value of the variable of interest, $X_{ij,;}$ and the coefficient from the Cox regression for that variable, β_i. (X_{ij} is the value of the ith variable X_i for the jth individual.)

(2) Calculate the hazard ratio for that variable for that individual at time $t_j = e^{(\beta_i * X_{ij})}$.
(3) Calculate the Schoenfeld residual as Y—hazard ratio.
(4) It is usually recommended to use weighted Schoenfeld residuals. To get the weighted Schoenfeld residual, divide the Schoenfeld residual from step 3 by the variance of β_i. The variance of β_i is the square of the standard error of β_i, i.e. (SE^2), which is provided in the statistical packages that do Cox regressions.
(5) Do this for every individual.
(6) Plot the residuals against time and see if they look randomly distributed around the zero line or if they show a pattern, like an upward or downward slope against time. If they look random, then the proportional hazards assumption has been met for that variable.
(7) Select times across the range of time in your study to make sure you have enough data points across the time span.

Another way to see if the assumption of a constant hazard ratio over time is violated is to include an interaction term for the variable by time, in your Cox model. If it is significant, then the assumption is violated.

One way to address a violation of proportional hazards is to stratify on the variable for which it is violated and do separate Cox regression analyses in each stratum. For example, say a particular variable has more effect in younger people than in older people. You might then stratify on age and do separate Cox regressions in the young and in the old people. If you use age as a continuous variable, you could get data points at different times across the time span of your study and use a variable in your model that is age at time t multiplied by time t.

Another way to address the proportional hazards violation is to include the variable as a time-dependent variable, i.e., the variable may change value as time goes on. For example, if you believe high blood pressure has more effect on dementia in middle age than at older ages, you might use blood pressure as a time-dependent variable. However, to do that you need multiple measures of blood pressure as time goes on. If you just have a baseline level of blood pressure, you can't make it time-dependent in your model.

Modeling is complex and requires deep knowledge of the field of study. It is important to consult experts in statistics as well as experts in the substantive domain under study.

4.21 Overlapping Confidence Intervals and Statistical Significance

Suppose you are looking at the risk of stroke in two groups, one being treated with a particular drug and the other with a different drug. You obtain a hazards ratio and 95% confidence interval for each group. Suppose further that the confidence intervals overlap. For example, in group A, you have an HR of 1.30 and 95% confidence

interval of 1.20–1.40, and in the other group, the confidence interval is 1.45–1.55. Can you conclude that the two groups are significantly different from each other with regard to risk of stroke? If the confidence intervals *do not* overlap, then the two groups are statistically significantly different from each other, and you can reject the hypothesis of no difference at the 0.05 level of significance. If these had been 99% confidence intervals, you could reject the null hypothesis of no difference at the 0.01 level of significance. However, if the confidence intervals *do* overlap, you cannot be sure that there is not a significant difference.

4.22 Confounding by Indication

One type of confounding that can occur in observational studies when you are looking at the effects of drug treatment on future events is *confounding by indication*. For example, suppose you want to compare the effects on heart disease of different drugs for high blood pressure. You determine what antihypertensive drugs study participants are taking at a baseline examination and then you follow them forward in time to see who develops heart disease. The problem may be that the reason people were taking different drugs at baseline was that they had different indications for them and the doctors prescribed medications appropriate to those indications. Thus, people with kidney disease may have been prescribed a different drug to control high blood pressure than those with angina or than those with no other medical conditions, and each of those indications may be differently related to the outcome of heart disease. Only a clinical trial, where the patients are randomly assigned to each drug treatment, can truly answer the question about different effects of the drugs.

However, there are ways to minimize the confounding by indication in observational studies; one way is to exclude from the analysis people who have angina or kidney disease in the above example. Another method gaining in use is *propensity analysis*.[24]

4.23 Propensity Analysis

The general idea behind propensity analysis is that you predict who is likely to be taking the drug from the independent variables you have measured and calculate an index of "propensity" for taking the drug. The propensity score is the probability of an individual having the exposure (taking the drug) conditional on his/her set of covariates. The propensity probability can be used in three different ways, and there is no consensus on which is best: (1) *regression adjustment*, (2) *stratification*, and (3) *matching*.

(1) *The first, called regression adjustment*, is to use the propensity score (or propensity probability) as a covariate in the logistic or Cox regression models (described in Section 4.20) that are looking at the relationship of your variable of interest to the outcome of interest, as described below. The propensity (or probability of exposure) is then entered as an independent variable in your final multivariate equation, along with a subset of the variables that you are controlling for. Each person's data then include the values of the covariates and his/her propensity score.

For example, in the hypertension example, if we are looking to see whether a calcium channel blocker is associated with mortality, we want to take into account that in an observational study, people might be more likely to have had a calcium channel blocker prescribed if they had angina, for example, and we know that angina is related to mortality. We might then take the following steps:

(a) Calculate a multiple logistic regression where Y (the dependent variable) = 1 if on drug, 0 otherwise

X_i (independent variables) = age, race/ethnicity, angina, BMI, systolic blood pressure, and other covariates that might influence prescribing a calcium channel blocker, like region of the country, socioeconomic indices, and so on. You should enter into this logistic as many baseline variables as you have measured to get a best fit model that predicts exposure.

(b) Calculate a propensity score for each person (probability of exposure based on the regression developed above).

(c) Calculate the regression you are really interested in which is to determine the association of calcium channel blockers with mortality after controlling for potential confounders, where Z = 1, if mortal event, 0 otherwise (dependent variable)

X_i = propensity score, age, race/ethnicity, plus some of the other relevant covariates that were in the original propensity equation.

(2) *The second way to use propensity scores is called stratification* (or subclassification). In this method, the propensity probabilities obtained from the logistic regression equation are divided into quintiles (or deciles), i.e., strata based on quintile of propensity. When subsequently obtaining the hazard ratio from a Cox regression equation (see Section 4.20), we would enter the drug of interest as an independent variable and the covariates that we wish to control for and stratify on propensity quintile (or decile). This means that the baseline hazard function is allowed to vary between the quintiles of propensity. What does that mean? Well, in each quintile stratum, the people are more or less equally likely to be on the drug (they have similar propensities). The hazard ratio (which is the hazard of the event in the group using the drug divided by the hazard in the group not using the drug, sometimes thought of as the relative risk) might be different in the group least likely to be using the drug (the bottom propensity quintile) than the hazard ratio in those most likely to use the drug

(those in the highest quintile). Stratification then compares people in groups similar in their propensity to use the drug.

Here is an example from a study of antidepressant use in the Women's Health Initiative.[25] The researchers were interested in the effects of antidepressant drugs on cardiovascular risk. The population studied consisted of 136,293 postmenopausal women who were followed for an average of 6 years. We will look at the effects of a particular antidepressant type of drug called selective serotonin reuptake inhibitors (SSRIs) on coronary heart disease and on stroke.

There are many factors that may influence a doctor prescribing an antidepressant, in general, and a given antidepressant, in particular. An especially important factor is the severity of the depression. Other factors may have to do with age, other comorbid conditions, region of the country (antidepressant use may be more common in certain geographic locations), and numerous other variables that could be potential confounders. The investigators wanted to know whether new use of antidepressants had an effect on subsequent heart disease or stroke risk. To address potential confounding by indication, the researchers obtained a propensity score from a logistic regression model to predict any new antidepressant use from 33 demographic, lifestyle, risk factor, and comorbidity variables measured at enrollment. (Some studies use upward of 100 variables for creating a propensity score, if these variables are available.) Thus, these propensity scores were a weighted composite of the individual covariates for each person.

To determine how good the logistic regression was at discriminating between those who used antidepressants versus those who didn't, we can use the c-statistic. The c-statistic is a measure of how much better than chance our prediction of drug use is based on the variables in our logistic regression model. A c-statistic of 0.5 would mean our logistic model is no better than chance at predicting membership in the drug group versus nonmembership. A c-statistic of 1.0 would mean the logistic regression is perfectly able to discriminate between people who use antidepressants versus those who don't. In our example, the c-statistic was 0.72, indicating a moderate ability of the variables included in the model to discriminate new use of antidepressants. (A c-statistic of 0.7 or higher is considered reasonable, and one of 0.8 or higher is considered to have strong discriminatory ability.)

The propensity scores were then divided into decile groups (quintiles are more commonly used and are perfectly adequate). The Cox regression model was then run using the STRATA statement in SAS software, version 9.1; (SAS Institute Inc., Cary, North Carolina). The dependent variable in the Cox regression was stroke, and the Cox model was stratified by decile of propensity to be taking any new antidepressant at the start of follow-up and adjusted for the following covariates: systolic blood pressure, body mass index, depression measure on a depression scale, hormone use, migraine or bad headache, aspirin or nonsteroidal anti-inflammatory use, and history of stroke or heart attack.

First, we look at the hazard ratio unadjusted for anything. For heart disease, which was defined as a heart attack or a death from heart disease, the unadjusted HR was 1.28 with 95% confidence intervals being 1.01–1.61 (usually written as 1.28

(95% CI: 1.01–1.61). This would lead us to believe that the new antidepressant users had a 28% greater risk of heart disease than nonusers. Since the confidence interval does not overlap 1.0, we are 95% confident that the true HR lies within that interval and thus that there is greater risk for users than nonusers. But, when we control for propensity to be on the drug and other covariates as described above, the hazard ratio is 0.95 (95% CI: 0.70–1.29), i.e., the confidence interval overlaps 1.0, and our conclusion is that the antidepressant is not associated with subsequent heart disease.

The situation was different for stroke, where the unadjusted HR was 1.40 (95% CI: 1.09–1.80) and the propensity adjusted HR was 1.45 (95% CI: 1.08–1.97)—indicating that risk of stroke was higher in users even after controlling for propensity and other covariates. Thus, it is very important to control for confounders.

A third way to use propensity scores is by matching. In this method, we select a person on the drug and find a person not taking the drug who has the closest propensity score; then, we take the next person on the drug and find that person's closest propensity match who is not on the drug and so on. Of course, there are computer programs that do that. The matching can be done up to five digits, four digits, and three digits. That is for the researcher to decide.

Thus, we form two groups: treated and control (matched on propensity), and that is the closest we can come to simulating randomization in an observational study. We can then compare these two groups on outcome. One problem is that we may not be able to find a match for each treated person, in which case the sample size (the matched groups) will be much less than we would have if we could match everyone and we may not have enough power to detect an effect (see chapter on power). In our example, the researchers were able to find only 4204 matched pairs or 8408 participants, out of the 136,293 women who were in the original analytic cohort. The two matched groups were similar on most baseline characteristics, indicating they were well balanced, except for several characteristics that were controlled for the in Cox models. The matched analysis provided a hazard ratio for stroke of 1.36 with 95% confidence intervals of 0.88–2.10. This compares to the 1.45 (95% CI: 1.08–1.97) found with propensity adjustment by stratification. The matched analysis still indicates the increased risk of drug use, but the confidence interval overlaps 1.0, reflecting the inadequate power because of the smaller sample size. But the order of magnitude of the hazard ratio is about the same.

In our example of antidepressants and cardiovascular disease, it is difficult to tease apart the effects of the antidepressants from the effects of severity of depression, since depression itself is a risk factor for cardiovascular disease. Due to problems of measurement of depression and variation in prescribing patterns, even propensity analysis may not be sufficient to disentangle these two variables.

Note several things about propensity scores. They can only control for known potential confounders, i.e., the variables that are available to the study. Unknown confounders of course cannot be controlled for. When you randomize people to a treatment and control group in a clinical trial, it is expected that the randomization will balance both known and unknown confounders in the two groups and thus avoid bias. And so the best way to determine the effects of antidepressants on outcomes

would be through controlled, randomized, double-blind clinical trials. This presents many challenges, including the cost, time, and feasibility of enrolling a large enough sample to have sufficient power to detect an effect. In the meantime, we may have to rely on observational study data, but we should remember that a randomized clinical trial is considered a "gold standard" of proof of causality. (See Chapter 6 on clinical trials.)

4.24 Selecting Variables for Multivariate Models

Suppose we want to determine the effect of depression on subsequent heart disease events using data from a prospective follow-up study. We can run Cox proportional hazards models to obtain relative risk of depression for heart disease endpoints, but we want to control for confounders. Otherwise, any association we see might really be a reflection of some other variable, like, say, smoking, which is related to depression and is also a risk factor for heart disease. How shall we go about deciding which variables to put in the model? There is no single answer to that question, and different experts hold somewhat different views, although it is generally agreed that known confounders should be included. So, we would put in variables that are significantly related to depression and also to heart disease among the nonexposed, i.e., nondepressed.

We would not include variables in the model that are presumed from past experience to be either highly correlated to depression (referred to as colinear) or intermediate in the pathway relating depression to heart disease, such as say some blood biomarker related to heart disease that is elevated by depression. In such a case, the elevation in the blood biomarker is intermediate between depression and heart disease; it may be the first manifestation of heart disease. We should not adjust for variables intermediate between exposure and outcome.

If we include variables related to the exposure but not the outcome, it will weaken the association of exposure and outcome. If we include variables related to outcome but not exposure, we may increase the precision of the estimate of the exposure–outcome relationship. If you control for variables related to the outcome but not the exposure, in general, it reduces the unexplained variation of the outcome variable and, therefore, reduces the variance of the parameter estimate for the exposure, whether it is a hazard ratio, odds ratio, or the coefficient in a linear regression. The point is that a lot of judgment has to be used in selecting variables for inclusion.

The objective is to see whether effects of depression that were found remain after accounting for other established risk factors. One strategy is to start by getting the hazard ratio of depression alone and then add successively, one at a time, other potential confounders to see if they change the hazard ratio for depression by 10% or more (though that is an arbitrary percentage). Variables that qualify by this criterion are kept in the model. For example, in the study of depression and deaths from cardiovascular causes, among postmenopausal women enrolled in the Women's Health Initiative, who had no prior cardiovascular disease, the hazard ratio

associated with depression controlling for age and race was 1.58; adding education and income to that resulted in a hazard ratio of 1.52. Adding additional variables to the model (diabetes, hypertension, smoking, high cholesterol requiring pills, hormone use, body mass index, and physical activity) didn't change things, resulting in a hazard ratio of 1.50. So, it was concluded that depression was an independent risk factor for cardiovascular death.

Now, if one were interested in developing a model that would predict risk (rather than one that would evaluate whether a particular risk factor was an independent contributor to risk, as in the example above), one might choose other strategies, like stepwise regression. Stepwise regression can be forward stepwise or backward stepwise, and computer programs calculating regressions ask you to specify which you want.

The basic principle is that in forward stepwise regression, you first enter the single variable that has the highest correlation with your outcome, then keep adding variables one at a time until you add one that is not statistically significant at some pre-chosen level, and then stop. In backward stepwise regression, you start out with all the potential variables that can be explanatory and drop them one at a time, eliminating the one that is least significant (has the highest p value) first, until dropping the next variable would result in a poorer model.

Many people don't like stepwise regression because it is somewhat arbitrary; it depends on the significance levels you chose to enter or leave the model, and also a variable may have quite a different effect if it is in a model with some other variables that might modify it, rather than when it is in the model alone. Another strategy is to look at all possible regressions—i.e., look at all two variable models, then at all possible three variable models, and so on. You select the best one according to how much of the variance in the dependent variable is explained by the model. An excellent discussion of variable selection in epidemiologic models is by Sander Greenland[26] and also in the advanced texts noted in the "Suggested Readings" section.

4.25 Interactions: Additive and Multiplicative Models

An interaction between two variables means that the effect of one variable on the outcome of interest is different depending on the level of the other variable, as described in Section 3.26. Interactions may be *additive*, where the joint effect of two variables is *greater than the sum of their individual effects*, or *multiplicative*, where the joint effect of the two variables *is greater than the product of the individual effects* of each variable.

Logistic and Cox regression models are inherently multiplicative. When we say that smoking carries a relative risk of 2 for coronary heart disease, for example, we mean that smokers are two *times* more likely than nonsmokers to get the disease. We may want to know if there is an interaction between smoking and hypertension. In other words, we want to know whether smoking among hypertensives has a greater

effect on heart attacks than we would expect from knowing the separate risks of smoking and hypertension. We can test the hypothesis of no interaction versus the alternative hypothesis of an interaction, but first we need to know what we would expect under a multiplicative model if there were no interactions.

Consider two dichotomous variables A (smoking) and B (hypertension).

The table below shows the pattern of relative risks expected under the multiplicative model if there really is no multiplicative interaction. The reference group is $RR_{no,no} = 1$. In other words, all our comparisons are to the risk among those who have neither A nor B, i.e., nonsmokers and nonhypertensives in our example. Note that $RR_{yes,yes} = RR_{yes,no} \times RR_{no,yes} = 2.0 \times 1.5 = 3.0$. ($RR_{yes,no}$ is the relative risk of B in the absence of A, and $RR_{no,yes}$ is the relative risk of A in the absence of B.)

Relative Risk (RR)

	A no	A yes
B no	$RR_{no,no} = 1$	$RR_{no,yes} = 1.5$
B yes	$RR_{yes,no} = 2.0$	$RR_{yes,yes} = 3.0$

If our observed $RR_{yes,yes}$ is significantly different from 3.0, we can reject the null hypothesis of no interaction and conclude that there is a multiplicative interaction.

To test this statistically, we would include a product term in our regression model, or Cox proportional hazards model (multiplying the value of variable B by the value of variable A for each person to get a new variable, which is the product of A and B), and then calculate the following quantity:

$$\frac{\widehat{\beta}(\textit{coefficient of the product term in the logistic regression})}{\textit{standard error of } \widehat{\beta}}$$

This quantity squared is approximately distributed as chi-square with 1 degrees of freedom.

What we are really testing is whether the coefficient β is significantly different from 0. If it is, then this is equivalent to concluding that the $RR_{yes,yes}$ is significantly different from our expected value of 3.0.

Note that if there is a significant interaction, we cannot interpret the main effects from that same model in the way we usually do. For example, if there were a significant interaction between smoking and hypertension, we would have to stratify on hypertension—i.e., look at the effects of smoking separately among hypertensives and nonhypertensives.

Suppose we wanted to see what kind of incidence figures might give rise to the table above. Remember, incidence is absolute risk, while relative risk is the absolute risk in one group relative to the absolute risk in the reference group. The two tables below both contain incidence figures that would give rise to the RR table above, so you can see it is possible to have different incidence rates or risks that have the same relative risk.

Risk or incidence per 1000

	A no	A yes
B no	$I_{no,no} = 20$	$I_{no,yes} = 30$
B yes	$I_{yes,no} = 40$	$I_{yes,yes} = 60$

	A no	A yes
B no	$I_{no,no} = 10$	$I_{no,yes} = 15$
B yes	$I_{yes,no} = 20$	$I_{yes,yes} = 30$

Additive risk is less commonly tested for, although some people think it should be. It is calculated from a difference in absolute risks (rather than from the ratio of absolute risks). Under the hypothesis of no interaction in an additive model, we would expect the data in the tables below.

Incidence per 1000

	A no	A yes
B no	$I_{no,no} = 20$	$I_{no,yes} = 30$
B yes	$I_{yes,no} = 40$	$I_{yes,yes} = 50$

Note: $\text{Incidence}_{yes,yes}$ = base incidence + effect of A + effect of B or

$$I_{yes,yes} = I_{no,no} + \left(I_{no,yes} - I_{no,no}\right) + \left(I_{yes,no} - I_{no,no}\right) = 20 + 10 + 20 = 50$$

Risk differences or attributable risk (AR) per 1000

	A no	A yes
B no	$AR_{no,no} = 0$	$AR_{no,yes} = 10$
B yes	$AR_{yes,no} = 20$	$AR_{yes,yes} = 30$

The $AR_{yes,yes}$ = effect of A plus effect of B = 10 + 20 = 30.

If the $AR_{yes,yes}$ is sufficiently different from the expected value of 30, then we may conclude there is an interaction on the additive scale.

The relative risk table that corresponds to the incidence table for the example given above of the additive model is:

Relative risk (RR)

	A no	A yes
B no	$RR_{no,no} = 1$	$RR_{no,yes} = 1.5$
B yes	$RR_{yes,no} = 2.0$	$RR_{yes,yes} = 2.5$

Thus, the expected value of $RR_{yes,yes}$ under the null hypothesis of no additive interaction is $RR_{yes,yes} = RR_{yes,no} + RR_{no,yes}$—1. If $RR_{yes,yes}$ is significantly different from the above expectation, we would be able to reject the null hypothesis of additive risk.

Interactions depend on the scale—i.e., whether we are talking about relative risks (multiplicative) or attributable risks (additive). It is wise to consult a statistician for appropriate interpretations.

Summary

Additive model interaction effect

See if observed value of $AR_{yes,yes}$ differs from expected value.

$$AR_{yes,yes} = AR_{yes,no} + AR_{no,yes}$$

or, $$RR_{yes,yes} = RR_{yes,no} + RR_{no,yes} - 1$$

Multiplicative model interaction effect

See if observed value of $RR_{yes,yes}$ differs from expected value.

$$RR_{yes,\ yes} = RR_{yes,\ no} \times RR_{no,\ yes}$$

4.26 Nonlinear Relationships: J Shape or U Shape

In the discussions of logistic regression and Cox proportional hazards models, we have been talking about linear relationships between the covariates in the model and the outcome or dependent variable. However, there are many biological phenomena that have a J- or U-shaped relationship where the risk of the outcome is higher at both the low end and high end of the exposure variable and lowest in the midrange. This is illustrated by Figure 4.5 showing the relationship of body mass index (BMI which is weight in kilograms divided by the square of the height in meters) to death among older men and women being treated for systolic hypertension in the Systolic Hypertension in the Elderly Program (SHEP).[27] The J or U shape means that as BMI goes up, the death rate goes up, but the death rate also goes up at very low values of BMI, hence the J or U shape. This may be due to preexisting illness: People who are on the very thin side may have lost weight because they are already ill and so of course they will be more likely to die. Or it may be due to the physiological consequences of very low weight.

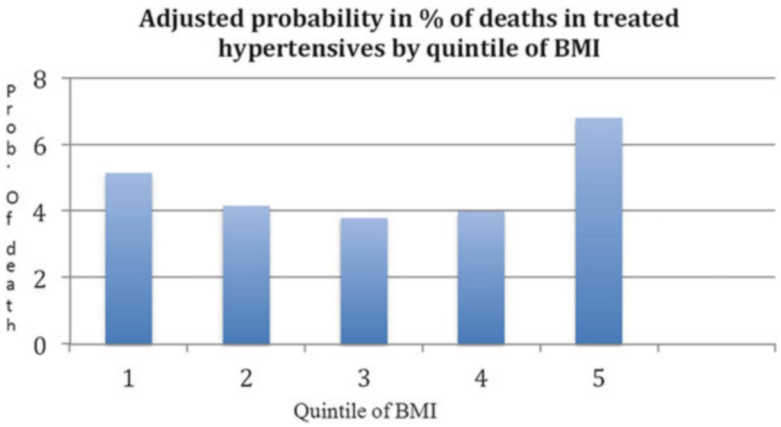


Figure 4.5 Adjusted probability in % of deaths in treated hypertensives by quintile of BMI

As you can see, the death rate per 100, adjusted for covariates, is higher in the lowest quintile of BMI and the highest quintile and lower in the second, third, and fourth quintiles.

To test whether there is a significant nonlinear relationship, we add the quadratic (square) term to the Cox proportional hazards model, so we would enter both BMI and BMI^2. (Of course first we would have to calculate BMI^2 for each person). In our example in this study, the coefficients we got, after controlling or adjusting for multiple variables, were for death:

	Coefficient	Standard error of the coefficient	*p* value
BMI	−0.3257	0.1229	0.008
BMI^2	0.005857	0.0020	0.003
	For stroke		
BMI	−0.2812	0.1482	0.06
BMI^2	0.0048	0.0024	0.05

Since the coefficient for BMI^2 is significant for both death and stroke, we conclude there is a nonlinear relationship between BMI and these outcomes. We cannot interpret the coefficients for BMI in the usual way because those reflect linear relationships, so they don't have the usual meaning now that we know the relationship is not linear.

From the Cox model, we can then calculate the hazard ratio or relative risk (we are using these interchangeably here) of death and stroke, for different values of BMI relative to the lowest value of BMI (the nadir of our curves), and we get the plots in Figure 4.6. Calculation of the nadir and the risks relative to the nadir are presented below the figure.

Figure 4.6 Adjusted relative risk of death and fatal and nonfatal stroke within the active treatment group by body mass index (BMI) (calculated as weight in kilograms

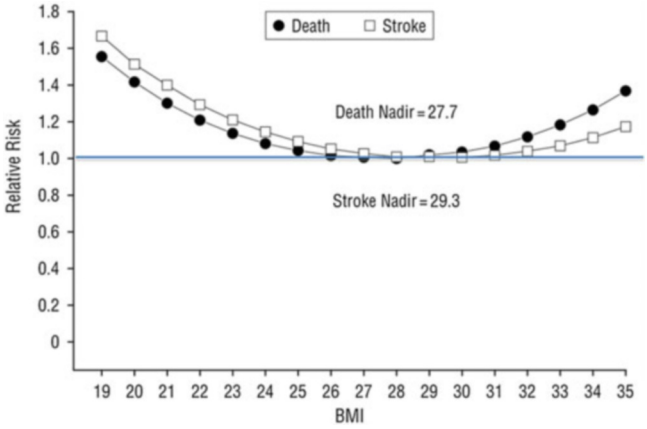


Figure 4.6 Relation of low body mass to death and stroke in the systolic hypertension in the elderly program. (Arch Intern Med. 2000; 160(4):494–500. doi:10.1001/archinte.160.4.494)

divided by the square of height in meters). The variables included educational level, history of diabetes, history of myocardial infarction, history of stroke, age, activity level, sex, race, smoking status, cholesterol level, BMI, BMI^2, systolic blood pressure (SBP), and SBP^2 (time dependent)

4.26.1 Nadir of Quadratic Relationship

If the quadratic (square term) coefficient is significant, we want to know at what level of BMI is the relative risk lowest, which is the nadir of the curve. The nadir for stroke is calculated as

$$\text{Nadir} = -\tfrac{1}{2}(\text{linear coefficient}/\text{quadratic coefficient})$$
$$\text{Nadir}_{\text{stroke}} = -\tfrac{1}{2}(-.2812/.0048) = -.5 \times -58.5833 = 29.3$$
$$\text{Nadir}_{\text{death}} = -\tfrac{1}{2}(-0.3257/0.005857) = 27.7$$

This means that the lowest mortality risk for this group occurred at a BMI of 27.7 and the lowest stroke risk occurred at a BMI of 29.3.

4.26.2 Risk Relative to Nadir

As an example, if we want to see what the risk of a BMI of 20 is relative to the risk of BMI of 29.3 (the nadir) for stroke, we can calculate it as shown below.

$RR = e^k$ (Sections 4.18 and 4.20)

$$k = \text{linear coefficient}(BMI_i - BMI_{nadir}) + \text{quadratic coefficient}\left(BMI_i{}^2 - BMI^2{}_{nadir}\right)$$
$$k = -.2812(20 - -29.3) + .0048\left(20^2 - 29.3^2\right)$$
$$= -.2812\ (-9.3) + .0048(-458.49)$$
$$= 2.6152 - 2.2008 = .4144$$
$$e^k = e^{.4144} = 1.51$$

Thus, compared to a BMI of 29.3, having a BMI of 20 is associated with a 51% higher risk of stroke after adjusting for multiple covariates. To get the points for Figure 4.6, we do such calculations for each BMI value for stroke and death separately, each with their own nadir.

4.26.3 Relative Risk Comparing Any Two Values

To calculate the relative risk of death for a given BMI value compared to any another BMI value,

$$k = \text{linear term coefficient} \times (BMI_1 - BMI_2) + \text{square term coefficient}\left(BMI_1{}^2 - BMI_2{}^2\right)$$
$$\left\{\text{Note: if the square term coefficient were not significant, then } k = \text{linear coefficient } [BMI_1 - BMI_2]\right\}$$

4.26.4 Confidence Intervals in U-Shaped Relationships for Risk Relative to a Specific Value

Consider the example of a U-shaped relationship between diastolic blood pressure (DBP) and death among older women in the Women's Health Initiative (WHI), where it was shown that both women with a very low diastolic blood pressure and women with a high diastolic blood pressure had higher risks of death than women whose blood pressure was near the nadir of 73 mmHg. A diastolic blood pressure of 80 mmHg or less is recommended by most hypertension guidelines. Suppose we want confidence intervals around the risk of a specific diastolic blood pressure compared to a DBP of 80 mmHg.

In a study in WHI of a subset of 6669 women, with average age of 79, followed prospectively for an average of 5 years, a Cox proportional hazards model, with the outcome of death, included a term for DBP and DBP^2 and controlled for multiple

other covariates. The quadratic term was significant, and the relevant coefficients are:

β_1 = coefficient of linear term = −0.10642
β_2 = coefficient of square term = 0.0007268
se (β_1) = standard error of β_1 = 0.02752
se (β_2) = standard error of β_2 = 0.0001852
Variance β_1 = se(β_1)^2 = 0.00075735
Variance β_2 = se(β_2)^2 = 0.000000034299
Covariance of (β_1, β_2) = −0.000005054
(Note that the nadir = − ½ (β_1/ β_2) = −0.5(−0.10642/0.0007268) = 73, but in this example we are not using the nadir as a reference point; instead, our comparison is a DBP of 80).
Thus, hazard ratio of DBP of 50 mmHg compared to a DBP of 80 mmHg = e^k.
where k = β_1 (DBP-80) + β_2 (DBP^2–80^2) = 0.35808,
and $e^k = e^{0.35808}$ = 1.43
Thus, a person with a DBP of 50 mmHg has a 43% higher risk of death than a person with a DBP of 80 mmHg.
Now we calculate the 95% CI for the hazard ratio = $e^{k-1.96\ se(k)}$, $e^{k\ +\ 1.96\ se(k)}$

To calculate the 95% CI for hazard ratio comparing two values, we need to compute the variance of k, so that we can get the standard error of k: se(k) = √variance(k).

Variance of k = (Var β_1) (DBP-80)^2 + (Var β_2)) (DBP^2–80^2)^2 + 2 Cov (β_1, β_2) (DBP-80) (DBP^2–80^2).

Thus, Var of k = 0.00075735 (50–80)^2 + 0.000000034299 (50^2–80^2)^2 + 2(−0.000005054)(50–80)(50^2–80^2) = 0.68161536 + 0.52168779– 1.182636 = 0.02066679.

se(k) = √variance(k) = √0.02066679 = 0.143762855.

Lower 95% confidence bound = $e^{k-1.96\ se(k)}$ = $e^{.35808-(1.96\ \times\ 0.14376)}$ = $e^{.07630}$ = 1.08.

Upper 95% confidence bound = $e^{k\ +\ 1.96\ se(k)}$ = $e^{.35808\ +\ (1.96\ \times\ 0.14376)}$ = $e^{.63986}$ = 1.90.

In summary, the risk of death associated with a DBP of 50 mmHg compared to a DBP of 80 mmHg is 1.43 with 95% confidence interval 1.08–1.90.

4.26.5 Restricted Cubic Splines

Another common way to fit regression models when the relationship between X and Y is nonlinear is to use restricted cubic splines. In this method you divide your exposure (x) into segments and fit a linear regression into each segment. The points of demarcation between segments are called knots. Thus, for example, you might have three knots, at say, the 25th, 50th, and 75th percentile of X. Then you would fit four regression lines: regression 1 for values up to the 25th percentile value of X,

regression 2 for X values between the 25th and 50th percentile of X, regression 3 for values between the 50th and 75th percentile of X, and regression 4 for value above the 75th percentile. If you have five knots, there would be six regressions. The computer programs that do this incorporate a method that ensures the fitted curves are smooth at the knots. Cubic splines are useful when the relationship between X and Y is more complex than a simple U shape or J shape. In the case of a U-shaped relationship, a regression that has a quadratic term in it, as described in the previous section, is sufficient and more interpretable.

4.27 Moderation of an Effect

First a note on terminology: we may be using the terms independent variable and predictor interchangeably and the terms dependent variable and outcome variable interchangeably.

An interaction between an independent variable X_1 and another independent variable X_2 means that the effect of X_1 on the outcome is different depending on the level of X_2. It means that the variable X_2 *moderates* (or modifies) the effect of X_1. And we say that X_2 is a *moderator variable*.

To determine if a variable is a moderator of another variable, you should test the interaction of these two variables. For example, if you are interested in the effect of smoking on heart disease and you suspect that this effect is greater in men than in women, then your outcome variable (Y) is heart attack, your predictor variable (X_1) is smoking, and your suspected moderating variable (X_2) is sex.

You can model this as follows:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \text{error}$$

If β_3 is significantly different from 0, then the interaction between sex and smoking means that the variable X_2 is a moderator of X_1.

If X_2 is sex, then you should stratify your analysis and look at men separately from women to see the effects of X_1 in women and the effects of X_1 in men:

$$Y = \alpha + \beta_{1m} X_1 + \text{error}$$

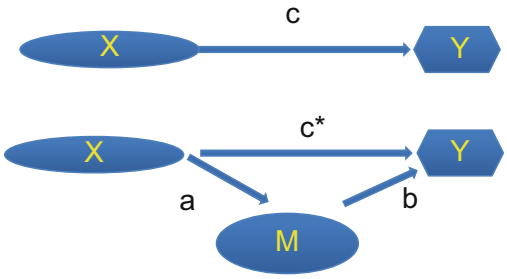
$$Y = \alpha + \beta_{1w} X_1 + \text{error}$$

$\beta_{1\text{men}}$ will be different from $\beta_{1\text{women}}$

4.28 Mediation of an Effect

Mediation is quite different. A mediating variable explains part of the effect of the independent variable of interest on the dependent or outcome variable. It is in the causal pathway between the independent variable X and the dependent variable Y. For example, the effect of a low dietary intake of potassium on increasing risk of stroke may be mediated by blood pressure. Dietary potassium reduces blood pressure, which in turn reduces stroke risk.

If the strength of the effect of an independent variable of interest is changed when you include the suspected mediator in the regression model, then that variable mediates the effect of your independent variable. Consider the diagrams below:



where X is the independent variable, M is the mediator variable, and Y is the outcome or dependent variable. The regression models that describe these relationships are shown below.

In the top part:

(1) $Y = i_c + cX + e_c$

c is the coefficient reflecting the total effect of X on Y; i is the intercept and e is the error term.

However, if there is mediation by a third variable, then part of the effect of X on Y is due to its effect on the mediator M (which in turn affects Y), and part is the direct effect of X on Y, as shown in the bottom part of the diagram.

(2) $M = i_a + aX + e_a$

a is the coefficient in the relationship of X (the independent variable) to M the mediator.

(3) $Y = i_b + bM + e_b$

b is the coefficient in the relationship of M (the mediator) to Y (the dependent variable).

(4) $Y_{c^*} = i_{c^*} + c^*X + b^*M + e$

c* is the coefficient in the relationship of X to Y when the mediator M is added to the model; b* is the coefficient in the relationship of X to Y when the X

variable is included in the model. (Note that the coefficient for X is denoted here as c* because it will be different from the coefficient c in Eq. 1, since now the new variable M is added to the model. Similarly, b* will be different from b in Eq. 3, because now the variable X is added to that model.)

It should be noted that other covariates may be added to these models, but if you do include other covariates, you must include them in all the models.

To determine that there is mediation, you must establish that:

(1) The independent variable X is related to the mediator M. This means that in Eq. 2, the coefficient a must be significantly different from zero.
(2) The mediator affects the outcome, after controlling for the independent variable. This means that in Eq. 4, the coefficient b* must be significant, i.e., non-zero.

The mediation effect is estimated to be the product term ab*.

If certain conditions hold, then in these special cases, it is equivalent to c-c*, or the difference between the effect of X on Y without the mediator, which is c, and the effect of X on Y when the mediator is accounted for, which is c*. The situation where ab* = c – c* requires that (1) the outcome variable is continuous, (2) multiple regression is used to estimate the coefficients, and (3) the same covariates are used in all the equations and that the same individuals are used in all the regression models.

Numerical example for the purpose of illustrating the methods:

Imagine a dataset for five people where the independent and moderator variables, X and M, and the outcome variable Y are as shown below.

X	M	Y
60	110	182
70	150	234
80	170	256
90	170	278
100	210	330

We calculate the following coefficients using linear regression:

(1) Y = cX = 3.4X—16.
(2) M = aX = 2.2 X—14.
(3) Y = c*X + b*M = 1.6X + 0.8182 M—4.5455.

(1) c = 3.4 is the total effect of x on y.
(2) a = 2.2 is the effect of X on the mediator.
(3) c* = 1.6 is the direct effect of X on Y, after accounting for the mediator.
(4) b* = 0.8182 is the effect of the mediator on Y, adjusting for X.

Thus, the strength of the relationship of X to Y is reduced from the total effect of 3.4 to 1.6 when the model includes the mediator M. The part of the relationship that is reduced, or mediated, is due to the relationship of the mediator to the outcome after adjusting for X.

The reduction in effect is c – c* = 3.4–1.6 = 1.8.

This is equivalent to ab* = 2.2 × 0.8182. = 1.8, which is the mediated effect.

The proportion mediated can be obtained by dividing the mediation effect by the total effect, and in the special case shown here, the three estimates below are equivalent:

ab*/c = 1.8/3.4 = 0.56.
or ab*/ (ab* + c*) = 1.8/((1.8 + 1.6)) = 0.56.
or c-c*/c = 1.8/((3.4)) = 0.56.

The equivalence of c-c* and ab* holds when the outcome variable is continuous and linear regression is used. Most commonly ab* is used as the estimate of the indirect or mediation effect. And the total effect is best calculated as the sum of the direct and indirect effect, or c* + ab*.

This is known as the "traditional approach" to mediation.

The method is the same if the moderator is dichotomous. However, if the outcome is dichotomous, then instead of linear regression, we use logistic regression, and the estimates c-c* versus ab* will generally not be equivalent.

4.28.1 Mediation with a Dichotomous Outcome: Logistic Regression

Numerical example:

In this example the outcome variable is dichotomous. Thus, we use logistic regression, rather than linear regression, to estimate the coefficients. When using logistic regression, the formulas are different for estimating direct and indirect effects, described below.

In a study in the Women's Health Initiative, it was found that a low dietary intake of potassium was associated with a higher risk of ischemic stroke. It is also known that higher potassium intake is associated with lower systolic blood pressure (SBP). The question is whether the effect of potassium on stroke risk is mediated by SBP. We considered potassium intake at the start of the study (at baseline) to see if it was related to ischemic stroke incidence and to see if SBP was the mediating factor.

We hypothesized that:low potassium ⟶ higher SBP ⟶ higher incidence of ischemic stroke.

In the example, the independent variable and the mediator are also dichotomous. The independent variable, or the exposure, is coded as low potassium intake if intake was less than 1925 mg daily; the mediator, systolic blood pressure, is coded as high if it was greater than 120 mmHg.

(1) First we look at path c in the diagram above to establish that the exposure (low potassium intake) is related to the outcome (ischemic stroke) by running a logistic regression analysis with ischemic stroke as the outcome and low potassium as the independent variable (X). We see that the coefficient of X, denoted by c, is significant. The data below indicate that low potassium intake is significantly associated with higher risk of stroke.

Independent variable	Intercept	Coefficient c	Standard error	p	Odds ratio	95% confidence limits
Low potassium	−3.7276	0.1483	0.0526	<0.0001	1.16	(1.05, 1.29)

Exposure is related to outcome.

(2) Next, we look at path a in the diagram to examine whether the independent variable is related to the mediator. We run a logistic regression analysis with the mediator (high SBP) as the outcome and low potassium as the independent variable. The result below shows that the coefficient a from the logistic regression is significant and we conclude that the mediator M (high SBP) is related to the exposure X (low potassium).

Independent variable	Intercept	Coefficient a	Standard error	p	Odds ratio	95% confidence limits
High SBP	0.3468	0.1119	0.0172	<0.0001	1.12	(1.08, 1.16)

Mediator is related to exposure

(3) Next we run a logistic regression analysis that includes both exposure and mediator and is represented by path c*. Probablity of stroke is a function of c* (low potassium) + b* (SBP).

Independent variable	Intercept	Coefficient	Standard error	p	Odds ratio	95% confidence limits
Low potassium —c*	−4.2208	0.1318	0.0527	<0.0001	1.14	(1.03, 1.26)
High SBP - b*		0.7403	0.0540	<0.0001	2.10	(1.89, 2.33)

Exposure is related to outcome after taking account of the mediator

You have now established that the exposure is related to the outcome, that the exposure is related to the mediator, that the mediator is related to the outcome, and that the effect of the exposure on the outcome is weaker when you account for the mediator. All this indicates that a mediation effect exists.

In the case where you assume there are no unobserved confounders (observed confounders can be controlled for), and you assume no interaction between the independent variable and the mediator, and when the outcome is dichotomous and the mediator is dichotomous, you can estimate direct and indirect effects through odds ratios.

The coefficients of interest are:

$i_a = 0.3468$
$a = 0.1119$
$b^* = 0.7403$
$c^* = 0.1318$

The odds ratio for the direct effect is $e^{c^*} = e^{.1318} = 1.14$.

The OR for the indirect effect is:

$$\frac{\left(1+e^{ia}\right)\left(1+e^{(b^*+ia+a)}\right)}{\left(1+e^{ia+a}\right)\left(1+e^{(b^*+ia\ )}\right)} = \frac{\left(1+e^{0.3468}\right)\ \left(\ 1+e^{(0.7403\ +\ 0.3468+\ 0.1119)}\right)}{\left(1+e^{(0.3468+\ 0.1119)\ \ (1+e(0.7401+\ 0.3468)}\right)} = 1.02$$

Total effect is $(e^{c^*}) \times (e^{ab^*}) = e^{0.1318} \times e^{.1119\ \times\ 0.7403} = 1.14 \times 1.09 = 1.24$.

(Note this is not the same as $e^{\bar{c}}$, which is $e^{.1483} = 1.16$ for a variety of reasons beyond the scope of this description. See the various articles in the references to this chapter for details. In general, when estimates of coefficients are made from logistic regressions, it is recommended that the total effect be estimated as the indirect effect plus the direct effect, i.e., on the odds ratio scale, it is $(e^{ab^*}) \times (e^{c^*})$ as in the formulas above. When coefficient estimates are on the log odds scale, then that is equivalent to sum of direct and indirect effects since $e^{(ab^*\ +\ c^*)} = (e^{ab^*}) \times (e^{c^*})$.

Proportion mediated is:

$$\frac{(\text{ OR direct})(\text{OR indirect}-1)}{(\text{OR direct})(\text{OR indirect})-1} = \frac{(1.14)(1.02-1)}{(1.14)(1.02)-1}$$

$$= \frac{.02}{.16} = 0.125 = 12.5\%$$

(Note: numbers in above calculations have been rounded to two decimal places.)

Another approach is called the potential outcomes approach, sometimes known as the counterfactual approach, which has a somewhat different terminology than that described above. For a dichotomous exposure, we say that $X = 1$ when the exposure is present and $X = 0$ when it is absent. We then ask what a person's Y or outcome value would be if his X value was 1 and what it would be if his X value was 0. We get the average of the Y values across all persons when $X = 1$ and the average of all persons if the value of X were 0 and the difference between these two averages is the effect of X on Y, when the mediator is not accounted for, or what we called c in diagram. It is the average Y, given $X = 1$ minus the average Y given $X = 0$ and is usually written as:

E[Y(1)] – E [Y(0)], where E means the expected value or average.

4.28.2 Controlled Direct Effect or CDE

This is the effect of X on Y, independent of the mediator, i.e., controlling for the mediator, as shown previously it is:

$$Y = i_{c^*} + c^*X + b^*M + e_{c^*}$$

4.28.3 Natural Direct Effect or NDE

Let us assume there may be an interaction between X and the mediator (as there often is) and also we may want to take into account other covariates, so the model for the direct effect becomes:

$$Y_{c^*} = i_{c^*} + c^*X + b^*M + d(XM) + (\text{other covariates}) + e_{c^*}$$

where d is the coefficient for the interaction term. In order to simplify, we will assume there are no other covariates we wish to control for.

$$\text{NDE} = c^* + d(i_a + aX) \text{ and the odds ratio for NDE is } e^{((c^* + d(ia + aX))}$$

When there is no interaction between X and M, then $d = 0$ and the natural direct effect and the controlled direct effect are the same, c*, and the odds ratio is e^{c^*}.

4.28.4 Natural Indirect Effect or NIE

This pertains to the effect of the mediator M on the outcome Y if the exposure X were controlled for.

$$\text{NIE} = ab^* + ad\ (X)$$

Again note that if there is no interaction, then $d = 0$ and the natural indirect effect is equal to the indirect (mediation) effect or ab*.

Different formulas and assumptions apply in different situations, depending on whether the outcome Y is continuous or discrete and the exposure X is continuous or discrete and the mediator M is continuous or discrete and to the various combinations of these possibilities.

Different estimation methods have been proposed to take into account possible exposure-independent variable interaction (XM interaction), or models that take into account other variables that may affect the mediator, or confounders or multiple

mediators, and so on. To add to the complexity, many papers on the subject use different symbols and it is necessary to translate one set of symbols to another when comparing articles. The point is that mediation analysis is a complex topic, still under development, and experts don't always agree in their recommendations on how to handle these various situations, but it is hoped that the above explanations will give the reader enough background to read papers that utilize mediation analysis and even to do simple analyses themselves when linear regression analyses can be used.

4.28.5 The Sobel Test for Significance of Mediation Effect

The mediation effect is estimated by the product of the two coefficients, when using linear regression methods: ab*.

To test the significance of the mediation, we may use the Sobel test:

$$Zab/SE_{ab^*}$$

SE is the standard error of ab*, which is the square root of the variance of ab*. SE_{ab^*} = square root [b^{*2} (SE of b*) + a^2 (SE of b*)] = ($b^{*2}\sigma_a{}^2 + a^2\sigma_{b^*}{}^2$).

However, this estimate of the standard error of ab* requires a large N, and there is another method of estimating SE_{ab^*} that is known as *bootstrapping*.

4.28.6 Bootstrapping

In bootstrapping you create a new sample dataset of the same size from your original dataset by randomly selecting data points from your original dataset; each data point in the original dataset has equal probability of being in the new sample dataset. So, for instance, say you start out with six numbers, 10, 9, 8, 7, 6, and 5, and you draw six numbers from that original dataset for a new sample. Each number in your original dataset would have 1/6 probability of being chosen for the new sample. So, your new sample might contain a 10, 10, 5, 5, 7, and 8. Now you draw sample 2 from the original dataset and you might come up with 9, 8, 8, 6, 7, and 9. You do this thousands and thousands of times (by a computer of course). It's basically sampling with replacement. Each of these new sample datasets of the original data has its own mean and standard deviation. You then get a distribution of those sample means and the standard deviation of that distribution of means is the estimate of the standard error of the original data. Now you can use that standard error in the Sobel test to test the mediation effect.

Another problem is that a and b* are assumed to be independent, but this assumption is often not true. An even more difficult problem is that when the

outcome is dichotomous and logistic regression is used, the indirect, or mediated, effect is not estimated by ab*, but rather by the formulas given in the above example from the Women's Health Initiative. How to calculate the standard error of those estimates is not clearly established.

4.29 What Is the Difference Between a Confounding Variable and a Mediating Variable?

A confounder is a variable that is related to both the exposure and the outcome, and a mediating variable is also related to both exposure and outcome so what is the difference between them? It is a theoretical difference and depends on the underlying model.

When a variable is a confounder, it means that the effect of the exposure on the outcome could really be due to the confounder. If there were perfect correlations between the exposure and the confounder, then you could substitute one for the other and you wouldn't know which was responsible for the association. A mediator, however, is an intermediate variable that is in the causal pathway. So the sequence of events is important here. The exposure causes a change in the mediator, which in turn causes a change in the outcome. So at least part of the relationship of the exposure to outcome is due to the effect of the exposure on the mediator. How you decide something is a confounder or a mediator depends on whether you are looking for a causal relationship or looking for an effect unmuddied by some other variable. It also depends on the timing of the associations. In a causal relationship, the mediator cannot occur before the independent variable, because the independent variable is causing a change in th mediator.

4.30 Penalized Regression: Lasso, Ridge, and Elastic Net Regression Methods

Suppose we want to predict an outcome, but we have very many variables that could be used in the prediction model and many of these variables are correlated with each other, (a situation known as multicollinearity). LASSO regression is one way we can select the most parsimonious set of those variables that will give us good prediction. Another method of handling a large set of variables is called ridge regression and a third method is elastic net regression. All these techniques are known as penalized regression because a penalty term is added to a quantity we want to minimize as part of the variable selection process, as will be explained below. All of these techniques are commonly used in machine learning algorithms.

4.30.1 LASSO Regression

LASSO stands for "Least Absolute Shrinkage and Selection Operator." It is also known as *L1 regularization*. The idea is that the LASSO procedure penalizes variables that have a weak relationship to the outcome or are considered not important, perhaps because they are correlated with other variables, by setting the beta coefficient for that variable close to zero or zero and thereby reducing the number of variables in the final selected set of predictors. Many computer statistical packages have the LASSO procedures, including SAS, STATA, and Python, R. This section is to provide an explanation of what the LASSO regressions do. In Lasso we try to get a balance between a simpler model (i.e., with fewer independent variables) and that still has good predictive ability. We do this as follows:

(1) First, we prepare the data. It is best to first convert our data to standardized variables. We do this for each variable by subtracting the mean of that variable from each value and dividing by the standard deviation.

It is what we previously called a Z-score and it is $= \dfrac{x_i\text{-mean}_x}{\text{standard dev.of x}}$

This then gives us a scaled variable with mean $= 0$ and sd $= 1$.

In this way, all the variables in our linear regression model will be on the same scale and the betas will be comparable for all variables. A beta coefficient for each variable will represent a unit change in the standardized variable.

(2) Now we consider a linear regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

where x_1 to x_k are all the potential predictor variables.

This is often referred to as ordinary least squares regression or OLS.

What we want is to make some of those β_i equal to zero so that we select only those variables that are important, so that we don't overfit the model. We want to make the model more sparse, but at the same time, we want the model to have a good fit, i.e., we want it to predict well.

One commonly used approach is to divide our original dataset into a training dataset and a test dataset. The training dataset is used to fit the regression model and determine the best parameters (i.e., the βetas), while the test dataset is used to see how well the model we developed on the training dataset works on new data. Often we can divide the data into 80% training and 20% test data. The training dataset can further be divided so that a part of it, say 20%, is used for validation and tuning of the parameters to further refine our choice of the best ones. Once we do that, we can then see how well our model, obtained in the training data, generalizes to new data in the test dataset.

To create and evaluate the model we came up with using the training dataset, we want to minimize the residual sum of squares (RSS) in the training dataset, which is a measure of the goodness of fit of the model.

The RSS = the sum across all individuals in the training dataset of the squared differences between predicted and observed values of the outcome Y:
Sum from j = 1 to N (all individuals) of $(Y_{observed} - Y_{predicted})^2 =$

$$\sum_{j=1}^{N} \left(Yj - \widehat{Y}j\right)2$$

(Note: predicted values are symbolized with $\hat{Y}$.)

(3) However, in order to shrink the model (i.e., eliminate some of the predictor variables), we will impose a penalty on the RSS that is a function of a parameter called lambda, λ, which is sometimes also known as alpha. The λ parameter controls the amount of shrinkage of the beta coefficients. We specify a value of λ we want to use and it can range from 0 to infinity. How to select a good λ is described in a later section, but the aim is to choose a λ that optimizes the final model for both parsimony and good predictive capacity. A bigger λ results in fewer predictors, i.e., a sparser model, but it increases the RSS, so we use an iterative process to balance out these two factors.

How the LASSO Algorithm Works to Shrink the Beta Coefficients

First let us consider the term "objective function." This refers to the objective of our minimization procedures and it means the function which it is the objective to minimize. In ordinary least squares (OLS) regression, we want to minimize the RSS for the best fit. In Lasso regression we want to minimize the RSS plus a penalty term. The objective function in LASSO regression, i.e., the quantity that LASSO regression tries to minimize, is:

RSS + $[\lambda (|\beta_1| + |\beta_2| + \ldots + |\beta_k|)]$, usually written as:

$$\sum_{j=1}^{N} \left(Yj - \widehat{Y}j\right)^2 + \lambda \sum_{i=1}^{k} |\beta i|,$$

where RSS is the residual sum of squares, $|\beta_i|$ is the absolute value of the ith beta coefficient, N is the total number of individuals, k is the number of variables (i.e., of the beta coefficients), Y_j is the actual outcome for the jth individual and $\widehat{Y}j$ is the predicted outcome for the jth individual.

Thus, the objective function is the residual sum of squares + λ times the sum of the absolute values of all the beta coefficients in the model, and $[\lambda * (|\beta_1| + |\beta_2| + \ldots + |\beta_k|)]$ is the penalty term. If λ (the amount of shrinkage) = 0, then there is no penalty and we are back to just plain regular regression, or OLS, where we want to minimize the residual sum of squares.

To shrink the beta coefficients and set some of them to zero, we start with the initial set of coefficients that we get from the ordinary linear regression using all the

predictor variables in the model. Then we use what is called *the coordinate descent algorithm*. In this approach we take one beta coefficient at a time, while keeping the rest of the βetas constant, and see what value of that one β_i will minimize the objective function, which is now called the "partial residual" and is the new RSS with the altered β_i and all the other betas at their original value.

To determine the new value of that specific β_i, we use what is called "soft-thresholding." The soft threshold for a particular β_i and a given λ (or regularization parameter) = the positive or negative sign of the β_i times whichever is higher (or the maximum of) either 0 or the absolute value of $(\beta_i - \lambda)$.

It is written as soft-threshold $(\beta_i \mid \lambda)$ = (+ or - sign of β_i) [maximum $(0, (|\beta_i| - \lambda)]$.

For example, if $\beta_i = \underline{0}.5$, and $\lambda = \underline{0}.1$, the soft threshold is + [max $(0, (\underline{0}.5 - \underline{0}.1)] = +0.4$.

The coefficient has been shrunk from $\underline{0}.5$ to $\underline{0}.4$.

If $\beta_i = -\underline{0}.5$, and $\lambda = \underline{0}.1$, the soft threshold is – [max $(0, (\underline{0}.5 - \underline{0}.1)] = -0.4$.

For another example, if $\beta_i = \underline{0}.08$, and $\lambda = 0.1$, the threshold is +[max $(0, (\underline{0}.08 - \underline{0}.1)] =$ [max $(0, -0.02)$].

The higher number between 0 and $-\underline{0}.02$ is 0, so the coefficient β_i is set to 0, and the predictor corresponding to that coefficient is eliminated from the model.

This process is iterated for each β in the model, one at a time, keeping all the other βetas constant (using the new values as they are developed) and calculating the partial residuals each time. The iteration stops after you reach a certain threshold for iteration, which could be a set number of iterations, or it could stop when the difference in the residual sum of squares differs from the last two iterations by a small amount.

Now the question is how do you pick a penalty parameter to use, i.e., how do you select λ? One common method is *cross-validation*, also known as *k-fold cross-validation*.

Cross-Validation to Select λ

In this method, first you select some lambda values you want to test, say you are considering four different lambda values. These are arbitrary, but you might want to test four values of different magnitude. Then you divide the training dataset into k subsets, also known as k folds. You might choose 5–20 folds depending on the size of your dataset. For this example, let's say you chose k to be 5. You hold out one of these folds (subsets of the training data) to use for validation, and train the data, i.e., fit the model, to the remaining k-1 folds using one of the four selected values of λ. You then test the model with that selected λ on the validation fold. Then you compute the mean square error (MSE), which is the residual sum of squares divided by N and is a measure of the performance of the model. Then you pick the next fold to hold out and train the new remaining k-1 folds using that same λ and again calculate the MSE. Thus you have five iterations for that λ. You average the MSE across these five iterations.

Now you go through the same process for the other three remaining values of λ, getting the average MSE for each lambda. You now have 20 iterations (5 folds by 4 lambda values) and 4 average MSE values, 1 for each lambda. You pick the λ that has the lowest average MSE to select your optimal lambda.

Next, you use the entire training dataset (all k folds together) to train your final model (i.e., to shrink your beta coefficients to get the more sparse model), using the optimal lambda you found through the cross-validation process. Finally, you evaluate that final model on the test dataset you created when you split the original full dataset into a training dataset and a test dataset (usually 20% of the full dataset that you partitioned off). This final step will tell your readers how well they can expect your LASSO model to work on some new dataset they might have.

LASSO in Logistic Regression

Recall that the logistic function is:

$$P(\text{event}) = P(Y = 1) = \frac{1}{1 + e^{-k}} = \frac{1}{1 + 1/e^{k}}$$

where $k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ (see Section 4.18).

When using LASSO for logistic regression, to evaluate how well our model performs we want to see how well the predicted probability of an event matches the actual occurrence of the event. So instead of wanting to minimize the RSS as we do in standard regression, we want to minimize the logistic objective function, also known as the cost function.

For each individual, j, we calculate that individual's cost function as:

$$-\left[Y_j \log\left(P_j\right) + \left(1 - Y_j\right) \log\left(1 - P_j\right)\right] + [\lambda \text{ sum } (|\beta_i|)]$$

where Y_j is 1 if the event occurred for that individual, or 0 if the event did not occur, and [λ sum ($|\beta_i|$)] is the penalty or regularization term.

Note if Y_j for an individual is 1 (i.e., if the event occurred for that individual), then that individual's cost function reduces to:

$$-\log\left(P_j\right) + [\lambda \text{ sum } (|\beta_i|)]$$

If Y_j for an individual is 0 (i.e., if the event did not occur for that individual), then that individual's cost function becomes:

$$-\log\left(1 - P_j\right) + [\lambda \text{ sum } (|\beta_i|)]$$

Now we sum these across all individuals, and add the penalty factor which is [λ sum ($|\beta_i|$)] and we have the objective function that we want to minimize when applied to logistic regression, which is:

$$-\sum\left[Y_j \ \log\left(P_j\right) + \left(1 - Y_j\right) \ \log\left(1 - P_j\right)\right] + \left[\lambda \sum |\beta_i|\right]$$

In words, it is the minus sign, times the sum across all individuals, of each individual's actual occurrence or non-occurrence of the event (Y_j) times the log of the predicted probability of the event (log P_j) plus 1 – occurrence of event ($1 - Y_j$) times log of 1 – predicted probability ($\log(1 - P_j)$), plus the penalty term, which is the regularization parameter λ times the sum of the absolute values of the coefficients for each predictor variable ($|\beta_i|$).

It should be noted that if you want to obtain hazard ratios or odds ratios and their confidence intervals, you can't just use the beta coefficients from the Lasso analysis because the standard errors of those coefficients are not readily available due to the bias introduced by the penalty term. There are several, rather complex, ways to estimate these, but perhaps the easiest thing to do would be to enter the variables selected by LASSO into a Cox or logistic regression and compute your hazard or odds ratios from those beta coefficients in the usual way. There are different approaches to various aspects of LASSO regression (like different optimization techniques or different ways of selecting the penalty factor), favored by different experts and used in different computer programs. It is hoped that the above explanations will be useful to the reader in making clearer the methodology sections of articles that use this procedure. However, if you are planning to do a LASSO analysis with your data, it is advisable to consult a statistician.

4.30.2 Ridge Regression

Ridge regression, also referred to as *L2 regularization*, is like LASSO regression (L1 regularization), except that the objective function we wish to minimize when we are shrinking the parameters (the betas) is different. In ridge regression, the penalty term is:

λ sum of β_i^2 while in LASSO it is:
λ sum of $|\beta_i|$.

Thus, in ridge regression, the objective function we want to minimize is:

$$\text{RSS + penalty} = \Sigma\,(Y_j - \hat{Y}_j)^2 + \lambda\Sigma\beta_i^2].$$

Ridge regression can shrink the βetas, but it doesn't set any of them to zero, as Lasso regression does, and thus it does not eliminate any variables from the model, although it can make some of them have very small effects. This is a useful procedure to deal with possible overfitting when you have a lot of variables, many

of which are correlated with each other (multicollinearity). Why is overfitting bad? Because if you use too many variables in your model, you might fit the training data very well, but then the model is so specific to the training dataset, that it won't predict as well in a new dataset. And remember we want a model will have good predictability in new data to which we will apply it.

4.30.3 Elastic Net Regression

Elastic net regression is a combination of LASSO and ridge regressions, since it uses both penalty terms. The objective function we want to minimize is:

$$\text{RSS} + \text{L1 penalty} + \text{L2 penalty} = \Sigma\left(Y_j - \hat{Y}_j\right)^2 + \lambda_1 \Sigma \mid \beta_i \mid + \lambda_2 \Sigma \beta_i^2.$$

It is particularly useful when you have a very great many variables and not so many individuals and also when there is a lot of multicollinearity.

4.30.4 Summary

The aim of these three methods is to create a model that can best predict the outcome of interest, where there are a large quantity of variables that are potential predictors (in fact sometimes more variables than people in the dataset), some of the variables are highly related to one another, and the aim is to ensure that the model that is developed on a training set of data will predict well on new, data unseen by the model. To do this, the models tune and refine the coefficients of the variables, eliminating or weakening some of them and making the model more sparse and interpretable. All this requires a lot of iteration and fine-tuning, involves a thorough knowledge of the actual data, and relies in some aspects on judgment calls. These new methods only became possible as computing power increased. Thus, LASSO was first introduced in 1996 by Robert Tibshirani, though ridge regression came first in 1970, developed by Hoerl and Kennard. Most recently, Zou and Hastie introduced elastic net regression in 2005. What is awe-inspiring is the creativity of the people who invented and pioneered these methods. Clearly statistics isn't static.

4.31 Meta-Analysis Versus Mega-Analysis

It is sometimes desirable to get a single estimate of an effect based on several or many different studies of the same question, perhaps because some studies are too small on their own or because the investigator wants to generalize more broadly than

to just one cohort. Both methods are a way of synthesizing data from multiple studies. In a *mega-analysis*, one pools all the individual-level data from each study cohort into one big (mega) dataset and proceeds to do the usual analyses on the large, pooled dataset. It is first necessary to harmonize the data because different studies may collect data on a given variable in different ways. This can be quite challenging. An example is Emerging Risk Factors Collaboration (ERFC) based in Cambridge UK, which has pooled data from about 2.5 million participants in more than 130 prospective studies from more than 30 countries to study risk factors for cardiovascular disease.

However, as is more often the case, individual-level data may not be available, but summary statistics are available, like means and standard deviations, for example, from which you can calculate effect size for each study. A meta-analysis pools such summary statistics, after weighting them by the size of the study and/or the precision of the effect size estimates, and arrives at a pooled effect size estimate[28].

Chapter 5
Mostly About Screening

I had rather take my chance that some traitors will escape detection than spread abroad a spirit of general suspicion and distrust, which accepts rumor and gossip in place of undismayed and unintimidated inquiry.

Judge Learned Hand (October 1952)

5.1 Sensitivity, Specificity, and Related Concepts

The issue in the use of screening or diagnostic tests is to strike the proper trade-off between the desire to detect the disease in people who really have it and the desire to avoid thinking you have detected it in people who really don't have it.

An important way to view diagnostic and screening tests is through sensitivity analysis. The definitions of relevant terms and symbols are as follows:

T+ means positive test, T– means negative test, D+ means having disease, and D– means not having disease. The symbol | means, "given that," so that P(T + | D–) means positive test, given that there is no disease or D–.

Sensitivity: the proportion of diseased persons the test classifies as positive,

$$= \frac{a}{a+c} = P(\mathrm{T}+|\mathrm{D}+); \quad \text{(probability of positive test, given disease)}$$

Specificity: the proportion of nondiseased persons the test classifies as negative,

$$= \frac{d}{b+d} = P(\mathrm{T}-|\mathrm{D}-); \quad \text{(probability of nagative test, given no disease)}$$

False-positive rate: the proportion of nondiseased persons the test classifies (incorrectly) as positive,

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_5

$$= \frac{b}{b+d} = P(\mathrm{T}+|\mathrm{D}-); \quad \text{(probability of positive test, given no disease)}$$

False-negative rate: the proportion of diseased people the test classifies (incorrectly) as negative,

$$= \frac{c}{a+c} = P(\mathrm{T}-|\mathrm{D}+); \quad \text{(probability of negative test, given disease)}$$

Predictive value of a positive test: the proportion of positive tests that identify diseased persons,

$$= \frac{a}{a+b} = P(\mathrm{D}+|\mathrm{T}+); \quad \text{(probability of disease given positive test)}$$

Predictive value of a negative test: the proportion of negative tests that correctly identifies nondiseased people,

$$= \frac{d}{c+d} = P(\mathrm{D}-|\mathrm{T}-); \quad \text{(probability of no disease given negative test)}$$

Accuracy of the test: the proportion of all tests that are correct classifications,

$$= \frac{a+d}{a+b+c+d}$$

Likelihood ratio of positive test: the ratio of probability of a positive test, given the disease, to the probability of a positive test, given no disease,

$$= \frac{P(\mathrm{T}+|\mathrm{D}+)}{P(\mathrm{T}+|\mathrm{D}-)} = \text{positive test, given disease versus positive test, given no disease}$$

$$= \frac{\text{sensitivity}}{\text{false positive rate}} = \frac{\text{sensitivity}}{1-\text{specificity}}$$

Likelihood ratio of a negative test:

$$= \frac{P(\mathrm{T}-|\mathrm{D}+)}{P(\mathrm{T}-|\mathrm{D}-)} = \text{negative test, given disease versus negative test, given no disease}$$

$$\frac{1-\text{specificity}}{\text{sensitivity}}.$$

Note also the following relationships:

(1) Specificity + the false-positive rate = 1:

$$\frac{d}{b+d}+\frac{b}{b+d}=1$$

therefore, if the specificity of a test is increased, the false-positive rate is decreased.

(2) Sensitivity + false-negative rate = 1:

$$\frac{a}{a+c}+\frac{c}{a+c}=1$$

therefore, if the sensitivity of a test is increased, the false-negative rate will be decreased.

Pretest probability of disease: The pretest probability of a disease is its prevalence. Knowing nothing about an individual and in the absence of a diagnostic test, the best guess of the probability that the patient has the disease is the prevalence of the disease.

Posttest probability of disease: After having the results of the test, the posttest probability of disease if the test is normal is $c/(c + d)$, and if it is abnormal, the posttest probability is $a/(a + b)$. The last is the same as the *predictive value of a positive test*.

A good diagnostic test is one that improves your guess about the patient's disease status over the guess you would make based on just the general prevalence of the disease. Of primary interest to a clinician, however, is the *predictive value of a positive test* (*PV+*), which is the proportion of people who have a positive test who really have the disease, $a/(a + b)$, and the *predictive value of a negative test* (*PV–*), which is the proportion of people with a negative test who really don't have the disease, $d/(c + d)$.

Sensitivity and specificity are characteristics of the test itself, but the predictive values are very much influenced by how common the disease is. For example, for a test with 95% sensitivity and 95% specificity used to diagnose a disease that occurs only in 1% of people (or 100 out of 10,000), we would have the following:

		Disease		
		Yes	No	
		+	–	
Test	+	95	495	590
	–	5	9,405	9,410
		100	9,900	10,000

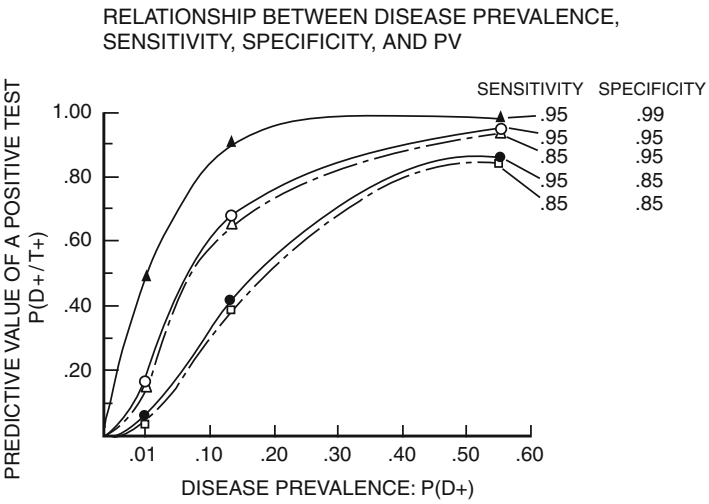


Figure 5.1 Relationship between prevalence, sensitivity, specificity, and predictive value

The PV+ is 95/590 = 0.16; that means that only 16% of all people with positive test results really have the disease; 84% do not have the disease even though the test is positive. The PV–, however, is 99.9%, meaning that if a patient has a negative test result, you can be almost completely certain that he really doesn't have the disease. The practical value of a diagnostic test is dependent on a combination of sensitivity, specificity, and disease prevalence, all of which determine the predictive values of test results.

If the prevalence of the disease is high, the predictive value of a positive test will also be high, but a good test should have a high predictive value even though the prevalence of the disease is low. Let us take a look at the relationship between disease prevalence and sensitivity, specificity, and predictive value of a test, shown in Figure 5.1.

Let us, for instance, consider a test that has a sensitivity of 0.95 and a specificity of 0.99. That means that this test will correctly label as diseased 95% of individuals with the disease and will correctly label as nondiseased 99% of individuals without the disease. Let us consider a disease whose prevalence is 10%; that is, 10% of the population have this disease, and let us now look and see what the predictive value of a positive test is. We note that it is approximately 0.90, which means that 90% of individuals with a positive test will have the disease. We can see that the predictive value of a positive test drops to approximately 0.70 for a test that has a sensitivity of 0.95 and a specificity of 0.95, and we can see that it further drops to approximately 0.40 for a test that has a sensitivity of 0.95 and a specificity of 0.85. In other words, only 40% of individuals with a positive test would truly have the disease for a test that has that particular sensitivity and specificity.

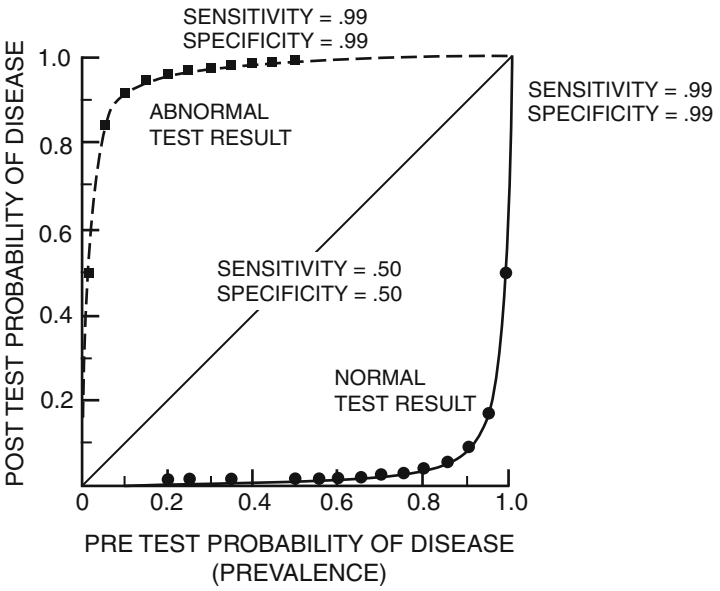


Figure 5.2 Posttest probability of disease versus prevalence for different sensitivities and specificities

One thing you can note immediately is that *for disease of low prevalence, the predictive value of a positive test goes down rather sharply*. The other thing that you can notice almost immediately is that large difference in sensitivity makes a small difference in the predictive value of a positive test and that a small difference in specificity makes a big difference in the predictive value of a positive test. This means that the characteristic of a screening test described by specificity is more important in determining the predictive value of a positive test than is sensitivity.

Figure 5.2 shows us a situation of a test that's virtually perfect. A test that has a sensitivity of 0.99 and a specificity of 0.99 is such that at most prevalence levels of disease, the probability of disease, given a normal or negative test result, is very low. That would be a very good test, and the closer we can get to that kind of situation, the better the diagnostic test is. The diagonal line in the center represents a test with a sensitivity of 0.50 and a specificity of 0.50, and that, of course, is a completely useless test because you can note that at the prevalence of the disease of 0.4, the probability of the disease given a positive test is also 0.4, which is the same as the probability of the disease without doing any test, and this pertains at each prevalence level. Therefore, such a test is completely useless, whereas a test with sensitivity and specificity of 0.99 is excellent, and anything in between represents different usefulness for tests. This, then, is an analytic way to look at diagnostic test procedures.

A particularly relevant example of the implications of prevalence on predictive value is the case of screening for the presence of infection with the AIDS virus. Since this disease is generally fatal, treatable but incurable at present, provokes high anxiety, has a stigma attached to it, and entails high costs, one would not like to

use a screening strategy that falsely labels people as positive for HIV, the AIDS virus.

Let us imagine that we have a test for this virus that has a sensitivity of 100% and a specificity of 99.995%, clearly a very, very good test. Suppose we apply it routinely to all female blood donors, in whom the prevalence is estimated to be very low, say 0.01%. In comparison, suppose we also apply it to homosexual men in San Francisco in whom the prevalence may be estimated to be 50% (this is for illustrative purposes only). For every 100,000 such people screened, we would have values as shown in the table on the following page.

Although in both groups all those who really had the disease would be identified, among female blood donors, one third of all people who tested positive would really not have the disease; among male homosexuals, only 6 out of 100,000 people with a positive test would be falsely labeled.

Positive Predictive Value as a Function of Prevalence

Test characteristics:

Sensitivity = 100%; Specificity = 99.995%; False positive rate = .005%

A. FEMALE BLOOD DONORS Prevalence = .01%,

True State

Screen Result	**HIV: +**	**HIV: –**	
+	10	5	15
–	0	99,985	99,985
	10	99,990	100,000

PV+ = 10/15 = .66667

B. MALE HOMOSEXUALS Prevalence = 50%,

True State

Screen Result	**HIV: +**	**HIV: –**	
+	50,000	3	50,003
–	0	49,997	49,997
	50,000	50,000	100,000

PV+ = 50,000/50,003 = .99994

5.2 Cutoff Point and Its Effects on Sensitivity and Specificity

We have been discussing sensitivity and specificity as characteristic of a diagnostic test; however, they can be modified by the choice of the *cutoff point between normal and abnormal*. For example, we may want to diagnose patients as hypertensive or normotensive by their diastolic blood pressure. Let us say that anyone with a diastolic pressure of 90 mmHg or more will be classified as "hypertensive." Since blood pressure is a continuous and variable characteristic, on any one measurement, a usually nonhypertensive individual may have a diastolic blood pressure of 90 mmHg or more, and similarly a truly hypertensive individual may have a single measure less than 90 mmHg. With a cutoff point of 90 mmHg, we will classify some nonhypertensive individuals as hypertensive, and these will be false positives. We will also label some hypertensive individuals as normotensive and these will be false negatives. If we had a more stringent cutoff point, say, 105 mmHg, we would classify fewer nonhypertensives as hypertensive since fewer normotensive individuals would have such a high reading (and have fewer false positives).

However, we would have more false negatives (i.e., more of our truly hypertensive people might register as having diastolic blood pressure less than 105 mmHg on any single occasion). These concepts are illustrated in Figure 5.3.

There are two population distributions, the diseased and nondiseased, and they overlap on the measure of interest, whether it is blood pressure, blood glucose, or other laboratory values. There are very few screening tests that have no overlap between normal and diseased individuals.

One objective in deciding on a cutoff point is to strike the proper balance between false positives and false negatives. As you can see in Figure 5.3, when the cutoff point is at A, all values to the right of A are called positive (patient is considered to have the disease). In fact, however, the patient with a value at the right of cutoff A

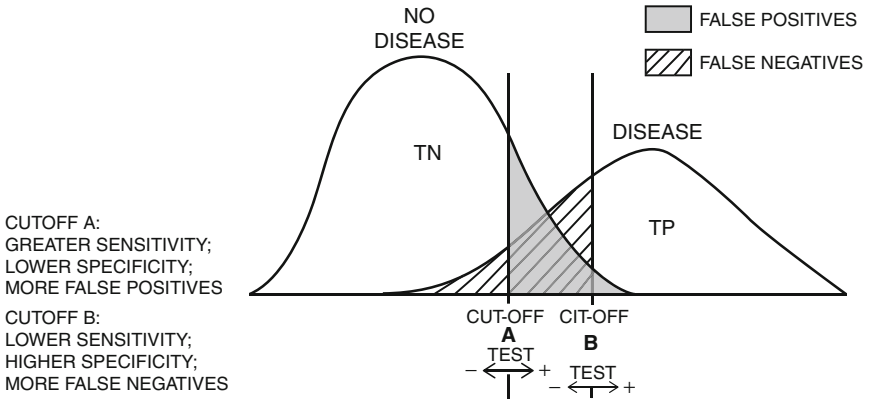


Figure 5.3 Different test cutoff points and false positives and false negatives

could come from the population of nondiseased people, since a proportion of people who are perfectly normal may still have values higher than those above A, as seen in the normal curve. The area to the right of A under the no-disease curve represents the false positive.

If an individual has a test value to the left of cutoff A, he may be a true negative or he may be a false negative because a proportion of individuals with the disease can still have values lower than cutoff A. The area under the "disease" curve to the left of cutoff A represents the proportion of false negatives.

If we move the cutoff point from A to B, we see that we decrease the area to the right of the cutoff, thereby decreasing the number of false positives but increasing the number of false negatives. Correspondingly, with cutoff A, we have a greater probability of identifying the truly diseased correctly, that is, pick up more true positives, thereby giving the test with cutoff A greater sensitivity. With cutoff B, we are less likely to pick up the true positives (lower sensitivity) but more likely to correctly identify the true negatives (higher specificity).

Thus, by shifting the cutoff point beyond what we call a test positive, we can change the sensitivity and specificity characteristics of the test. The choice of cutoff, unless there is some special physiological reason, may be based on consideration of the relative consequences of having too many false positives or too many false negatives. In a screening test for cancer, for example, it would be desirable to have a test of high sensitivity (and few false negatives), since failure to detect this condition early is often fatal. In a mass screening test for a less serious condition or for one where early detection is not critical, it may be more desirable to have a high specificity in order not to overburden the health-care delivery system with too many false positives. Cost consideration may also enter into the choice of cutoff point.

The relationship between sensitivity (the ability to correctly identify the diseased individuals) and the false-positive fractions is shown in Figure 5.4.

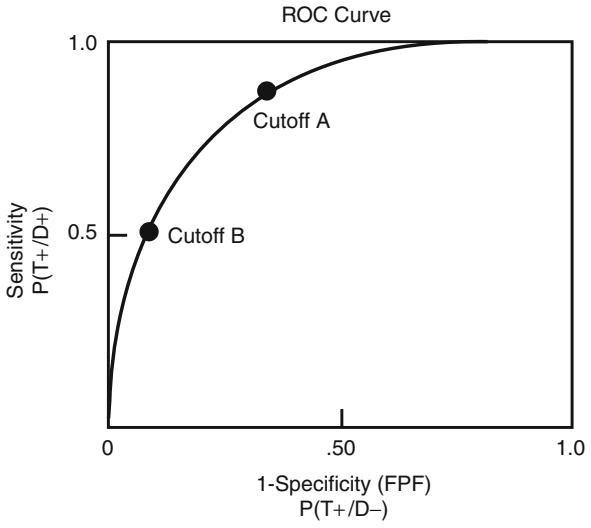


Figure 5.4 Receiver operating characteristic (ROC) curve, sensitivity versus false-positive fraction

This is called the receiver operating characteristic (ROC) curve of the test. Often we can select the cutoff point between normal and abnormal, depending on the trade-off we are willing to make between sensitivity and the proportion of false positives.

We can see that with cutoff A, while we can detect a greater percentage of truly diseased individuals, we will also have a greater proportion of false-positive results, while with cutoff B, we will have fewer false positives but will be less likely to detect the truly diseased. Screening tests should have corresponding ROC curves drawn.

Chapter 6
Mostly About Clinical Trials

It is no easy task to pitch one's way from truth to truth through besetting errors.

Peter Marc Latham (1789–1875)

I wouldn't have seen it if I didn't believe it!

Attributed to Yogi Berra

Unfortunately, sometimes scientists see what they believe instead of believing what they see. Randomized, controlled clinical trials are intended to avoid that, and other kinds, of bias.

A *randomized clinical trial* is a prospective experiment to compare one or more interventions against a control group in order to determine the effectiveness of the interventions. A clinical trial may compare the value of a drug versus a placebo. A placebo is an inert substance that looks like the drug being tested. It may compare a new therapy with a currently standard therapy, surgical with medical intervention, two methods of teaching reading, and two methods of psychotherapy. The principles apply to any situation in which the issue of who is exposed to which condition is under the control of the experimenter and the method of assignment is through randomization.

6.1 Features of Randomized Clinical Trials

(1) There is a group of patients who are designated study patients. All criteria must be set forth and met before a potential candidate can be considered eligible for the study. Any exclusions must be specified.
(2) Any reasons for excluding a potential patient from participating in the trial must be specified prior to starting the study. Otherwise, unintentional bias may enter. For example, suppose you are comparing coronary bypass surgery with the use of a new drug for the treatment of coronary artery disease. Suppose a patient comes along who is eligible for the study and gets assigned to the surgical treatment. Suppose you now discover the patient has kidney disease. You decide

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_6

to exclude him from the study because you think he may not survive the surgery with damaged kidneys. If you end up systematically excluding all the sicker patients from the surgical treatment, you may bias the results in favor of the healthier patients, who have a better chance of survival in any case. In this example, kidney disease should be an *exclusion criterion applied to the patients before they are assigned to any treatment group.*

(3) Once a patient is eligible, he or she is randomly assigned to the experimental or control group. Random assignment is not "haphazard" assignment, but rather it means that each person has an equal chance of being an experimental or control patient. It is usually accomplished by the use of a table of random numbers, described later, or by computer-generated random numbers.

(4) Clinical trials may be double-blind, in which neither the treating physician nor the patient knows whether the patient is getting the experimental treatment or the placebo; they may be single-blind, in which the treating physician knows which group the patient is in but the patient does not know. A double-blind study contains the least bias but sometimes is not possible to do for ethical or practical reasons. For example, the doctor may need to know the group to which the patient belongs so that medication may be adjusted for the welfare of the patient. There are also trials in which both patients and physicians know the treatment group, as in trials comparing radical mastectomy versus lumpectomy for treatment of breast cancer. When mortality is the outcome, the possible bias introduced is minimal, provided that exclusion criteria were specified and applied before eligibility was finally determined and that the randomization of eligible participants to treatment groups was appropriately done.

(5) While clinical trials often compare a drug or treatment with placebo, they may also compare two treatments with each other or a treatment and "usual care." Trials that compare an intervention with "usual care" obviously cannot be blinded, for example, comparing a weight-loss nutritional intervention with "usual" diet; however, the assessment of effect (measurement of weight, or blood pressure, or some hypothesized effect of weight loss) should be done in a blinded fashion, with the assessor not knowing which group the participant has been assigned to.

(6) It is essential that the control group be as similar to the treatment group as possible so that differences in outcome can be attributed to differences in treatment and not to different characteristics of the two groups. Randomization helps to achieve this comparability.

(7) We are concerned here with Phase III trials. New drugs have to undergo Phase I and II trials, which determine toxicity, and safety and efficacy, respectively. These studies are done on small numbers of volunteers. Phase III trials are large clinical trials, large enough to provide an answer to the question of whether the drug tested is better than placebo or than a comparison drug.

6.2 Purposes of Randomization

The basic principle in designing clinical trials or any scientific investigation is to *avoid systematic bias*. When it is not known which variables may affect the outcome of an experiment, the best way to avoid systematic bias is to assign individuals into groups randomly. *Randomization* is intended to ensure an approximately equal distribution of variables among the various groups of individuals being studied. For instance, if you are studying the effect of an antidiabetic drug and you know that cardiac risk factors affect mortality among diabetics, you would not want all the patients in the control group to have heart disease, since that would clearly bias the results. By assigning patients randomly to the drug and the control group, you can expect that the distribution of patients with cardiac problems will be comparable in the two groups. Since there are many variables that are unknown but may have a bearing on the results, randomization is insurance against unknown and unintentional bias. Of course, when dealing with variables known to be relevant, one can take these into account by *stratifying and then randomizing within the strata*. For instance, age is a variable relevant to diabetes outcome. To stratify by age, you might select four age groups for your study: 35–44, 45–54, 55–64, and 65 plus. Each group is considered a stratum. When a patient enters into the clinical trial, his age stratum is first determined, and then he is randomly assigned to either experimental or control groups. Sex is another variable that is often handled by stratification.

Another purpose of randomization has to do with the fact that the statistical techniques used to compare results among the groups of patients under study are valid under certain assumptions arising out of randomization. The mathematical reasons for this can be found in the more advanced texts listed in the Suggested Readings.

It should be remembered that sometimes randomization fails to result in comparable groups due to chance. This can present a major problem in the interpretation of results, since differences in outcome may reflect differences in the composition of the groups on baseline characteristics rather than the effect of intervention. Statistical methods are available to adjust for baseline characteristics that are known to be related to outcome. Some of these methods are logistic regression, Cox proportional hazards models, and multiple regression analyses.

6.3 How to Perform Randomized Assignment

Random assignment into an experimental group or a control group means that each eligible individual has an equal chance of being in each of the two groups. This is often accomplished by the use of random number tables. For example, an excerpt from such a table is shown below:

48461	70436	04282
76537	59584	69173

Its use might be as follows. All even-numbered persons are assigned to the treatment group, and all odd-numbered persons are assigned to the control groups. The first person to enter the study is given the first number in the list, the next person gets the next number, and so on. Thus, the first person is given number 48461, which is an odd number and assigns the patient to the control group. The next person is given 76537; this is also an odd number so he/she too belongs to the control group. The next three people to enter the study all have even numbers, and they are in the experimental group. In the long run, there will be an equal number of patients in each of the two groups.

6.4 Two-Tailed Tests Versus One-Tailed Test

A clinical trial is designed to test a particular hypothesis. One often sees this phrase in research articles: "Significant at the 0.05 level, two-tailed test." Recall that in a previous section, we discussed the concept of the "null hypothesis," which states that there is no difference between two groups on a measure of interest. We said that in order to test this hypothesis, we would gather data so that we could decide whether we should reject the hypothesis of no difference in favor of some alternate hypothesis. *A two-tailed test* versus *a one-tailed test refers to the alternate hypothesis posed.* For example, suppose you are interested in comparing the mean cholesterol level of a group treated with a cholesterol-lowering drug to the mean of a control group given a placebo. You would collect the appropriate data from a well-designed study, and you would set up the null hypothesis as.

H_o:	Mean cholesterol in treated group = mean cholesterol in control group
You may choose as the alternate hypothesis	
H_A:	Mean cholesterol in treated group is *greater than* the mean in controls

Under this circumstance, you would reject the null hypothesis in favor of the alternate hypothesis if the observed mean for women was sufficiently *greater* than the observed mean for men, to lead you to the conclusion that such a great difference in that direction is not likely to have occurred by chance alone. This, then, would be a one-tailed test of the null hypothesis.

If, however, your alternate hypothesis was that the mean cholesterol level for females is *different* from the mean cholesterol level for males, then you would reject the null hypothesis in favor of the alternate *either* if the mean for women was *sufficiently greater* than the mean for men *or* if the mean for women was *sufficiently lower* than the mean for men. The direction of the difference is not specified. In medical research, it is more common to use a two-tailed test of significance since we often do not know in which direction a difference may turn out to be, even though we may think we know before we start the experiment. In any case, it is important to report whether we are using a one-tailed or a two-tailed test.

6.5 Clinical Trial as "Gold Standard"

Sometimes observational study evidence can lead to misleading conclusions about the efficacy or safety of a treatment, only to be overturned by clinical trial evidence, with enormous public health implications. The Women's Health Initiative (WHI) clinical trial of hormone therapy is a dramatic example of that.[22] Estrogen was approved by the FDA for relief of postmenopausal symptoms in 1942, aggressively marketed in the mid-1960s, and after 1980, generally combined with progestin for women with a uterus because it was found that progestin offset the risks of estrogen for uterine cancer. In the meantime, many large prospective follow-up studies almost uniformly showed that estrogen reduced heart diseases by 30–50%. In the 1993, WHI mounted a large clinical trial to really answer the question of long-term risks and benefits of hormone therapy. One part was the study of estrogen alone for women had had a hysterectomy and thus didn't need progestin to protect their uterus, and another part was of estrogen plus progestin (E + P) for women with an intact uterus.

The E + P trial was a randomized, double-blind, placebo-controlled clinical trial meant to run for an average of 8.5 years. It included 16,608 women ages 50–79; such a large sample size was deemed necessary to obtain adequate power. The trial was stopped in 2002, 3 years before its planned completion, because the Data and Safety Monitoring Board or DSMB (as described in Chapter 10) found estrogen plus progestin caused an excess of breast cancer, and surprisingly, there was a significant and entirely unexpected excess of heart attacks in the E + P group compared to placebo! Final results, reported in subsequent papers, showed that the adverse effects (a 24% increase in invasive breast cancer, 31% increase in strokes, 29% increase in coronary heart disease, and more than a twofold increase in pulmonary embolism and in dementia) offset the benefits (a 37% decrease in colorectal cancer and 34% decrease in hip fractures), so that taken together, the number of excess harmful events per year was substantial. Since there were six million women taking this preparation in the United States alone, and millions more globally, these results have important implications for women other than those in the trial itself.

Why such different results from a clinical trial than from observational longitudinal studies? The most likely explanation is selection bias. Women who were taking hormones and then followed to observe their rates of heart disease were, in virtually all the observational studies, healthier, thinner, more active, more educated, and less overweight than their non-hormone-taking counterparts, and their healthier lifestyle and better baseline health status, rather than the hormones per se, was what accounted for their lower rates of heart disease.

The question now is answered using the "gold standard," the clinical trial: Estrogen plus progestin does not protect against heart disease and in fact increases the risk. As noted before, the impact of this research is great since so many millions of women were using the preparation tested.

6.6 Regression Toward the Mean

When you select from a population those individuals who have high blood pressure and then at a later time measure their blood pressure again, the average of the second measurements will tend to be lower than the average of the first measurements and will be closer to the mean of the original population from which these individuals were drawn. If between the first and second measurements you have instituted some treatment, you may incorrectly attribute the decline of average blood pressure in the group to the effects of treatment, whereas part of that decline may be due to the phenomenon called *regression toward the mean*. (That is one reason why a placebo control group is most important for comparison of effects of treatment above and beyond that is caused by regression to the mean.) Regression to the mean occurs when you select out a group because individuals have values that fall above some criterion level, as in screening. It is due to variability of measurement error. Consider blood pressure.

The observed value of blood pressure is the person's true value plus some unknown amount of error. The assumption is that people's measured blood pressure is normally distributed around the mean of their true but unknown value of blood pressure. Suppose we will only take people into our study if their blood pressure is 160 or more. Now suppose someone's true systolic blood pressure is 150, but we measure it 160. We select that person for our study group just because his measured value is high. However, the next time we measure his blood pressure, he is likely to be closer to his true value of 150 than the first time. (If he had been close to his true value of 150 the first time, we would never have selected him for our study to begin with, since he would have been below our cutoff point. So he must have had a large error at that first measurement.) Since these errors are normally distributed around his true mean of 150, the next time we are more likely to get a lower error and thus a lower measured blood pressure than the 160 that caused us to select him/her.

Suppose now that we select an entire subgroup of people who have high values. The averages of the second measurements of these selected people will tend to be lower than the average of their first measurements and closer to the average of the entire group from which we selected them. The point is that people who have the highest values the first time do not always have the highest values the second time because the correlation between the first and second measurement is not perfect. Similarly, if we select out a group of people because of low values on some characteristic, the average of the second measurements on these people will be higher than the average of their first measurements and again closer to the mean of the whole group.

Another explanation of this phenomenon may be illustrated by the following example of tossing a die. Imagine that you toss a die 360 times. Whenever the die lands on a five or a six, you will toss the die again. We are interested in three different averages: (1) the mean of the first 360 tosses, (2) the mean of the tosses that will result in our tossing again, and (3) the mean of the second tosses. Our results are shown in the table below.

Although on the first toss the mean of the 360 tosses is 3.5, we only pick the two highest numbers, and *their* mean is 5.5. These 120 times when the die landed on 5 or 6 will cause us to toss again, but on the second toss, the result can freely vary between 1 and 6. Therefore, the mean of the second toss must be lower than the mean of the group we selected from the first toss specifically because it had the high values.

First toss		Second toss	
Result	# of times result is obtained	Result	# of times result is obtained
1	60		
2	60		
3	60		
4	60		
5	*60*	1	20
6	*60*	2	20
		3	20
		4	20
		5	20
Mean of 360 tosses = 3.5		6	20
Mean of the 120 tosses that landed 5 or 6 = 5.5		Mean of the second toss = 3.5	

6.7 Intention-to-Treat Analysis

Data from clinical trials in general should be analyzed by comparing the groups as they were originally randomized and not by comparing to the placebo control group only those in the drug group who actually did take the drug. The people assigned to the active drug group should be included with that group for analysis even if they never took the drug. This may sound strange, since how can one assess the efficacy of a drug if the patient isn't taking it? But the very reason people may not comply with the drug regimen may have to do with adverse effects of the drug, so that if we select out only those who do comply, we have a different group from the one randomized, and we may have a biased picture of the drug effects.

Another aspect is that there may be some quality of compliers in general that affects outcome. A famous example of misleading conclusions that could arise from not doing an intention-to-treat analysis comes from the Coronary Drug Project. This randomized, double-blind study compared the drug clofibrate to placebo for reducing cholesterol. The outcome variable, which was 5-year mortality, was very similar in both groups, 18% in the drug group and 19% in the placebo group. It turned out, however, that only about two thirds of the patients who were supposed to take clofibrate actually were compliant and did take their medication. These people had a 15% mortality rate, significantly lower than the 19% mortality in the placebo group. However, further analysis showed that among those assigned to the placebo group,

one third didn't take their placebo pills either. The two thirds of the placebo group who were compliant had a mortality of 15%, just like the ones who complied with the clofibrate drug! The noncompliant people in both the drug and placebo groups had a higher mortality (25% for clofibrate and 28% for placebo). It may be desirable in some circumstances to look at the effect of a drug in those who actually take it. In that case, the comparison of drug compliers should be to placebo compliers rather than to the placebo group as a whole.

The inclusion of non-compliers in the analysis dilutes the effects, so every effort should be made to minimize noncompliance. In some trials, a judged capacity for compliance is an enrollment criterion, and an evaluation is made of every patient as part of determining his or her eligibility as to whether this patient is likely to adhere to the regimen. Those not likely to do so are excluded prior to randomization. However, if the question at hand is how acceptable is the treatment to the patient, in addition to its efficacy, then the basis for inclusion may be the general population who might benefit from the drug, including the non-compliers.

In the Women's Health Initiative, the primary analysis was intention to treat. However, a secondary analysis adjusted for compliance (more commonly referred to as adherence). In this analysis, the event history of the participant was censored 6 months after she either stopped taking the study pills or was taking less than 80% of the study pills. In the placebo group, the event history was censored 6 months after the participant started taking hormones (some participants in the placebo group stopped taking study pills but were prescribed hormones by their physicians and started taking them on their own). Thus, this secondary analysis basically compared the two groups "as treated" rather than as assigned to a particular treatment. In the intention-to-treat analysis, the hazard ratio for coronary heart disease was 1.24, while in the "adherence-adjusted" analysis, it was 1.50. Thus, the findings from the intention-to-treat analysis were confirmed and strengthened in the adherence-adjusted analyses.

6.8 How Large Should the Clinical Trial Be?

A clinical trial should be large enough, that is, have big enough sample size, to have a high likelihood of detecting a *true* difference between the two groups. If you do a small trial and find no significant difference, you have gained no new information; you may not have found a difference simply because you didn't have enough people in the study. You cannot make the statement that there is no difference between the treatments. If you have a large trial and find no significant difference, then you are able to say with more certainty that the treatments are really not different.

Suppose you do find a significant difference in a small trial with $p < 0.05$ (level of significance). This means that the result you obtained is likely to have arisen purely by chance less than 5 times in 100 (if there really were no difference). Is it to be trusted as much as the same p value from a large trial? There are several schools of thought about this.

The *p* value is an index of the strength of the evidence with regard to rejecting a null hypothesis. Some people think that a *p* value is a *p* value and carries the same weight regardless of whether it comes from a large or small study. Others believe that if you get a significant result in a small trial, it means that the effect (or the difference between two population means) must be large enough so that you were able to detect it even with your small samples, and therefore, it is a meaningful difference. It is true that if the sample size is large enough, we may find statistical significance if the real difference between means is very, very small and practically irrelevant. Therefore, finding a significant difference in a small trial does mean that the effect was relatively large.

Still others say that in practice, however, you can have less confidence that the treatments really do differ for a given *p* value in a small trial than if you had obtained the same *p* value in testing these two treatments in a large trial.[29] This apparent paradox may arise in situations where there are many more small trials being carried out worldwide studying the same issue than there are large trials—such as in cancer therapy. Some of those trials, by chance alone, will turn out to have significant results that may be misleading.

Suppose that there are 1000 small trials of anticancer drug therapy. By chance alone, 5% of these will be significant even if the therapies have no effect, or 50 significant results. Since these are by chance alone, it means we are incorrect to declare anticancer drug effects in these trials (we have committed type I errors). Suppose, further, that there are only 100 large trials studying this same issue. Of these, 5%, or five such studies, will declare a difference to exist, incorrectly. So if we combine all the trials that show significant differences *incorrectly*, we have 55 such significant but misleading *p* values. Of these, 50% or 91% come from small trials and 5 out of the 55 incorrect ones (or 9%) come from the large trials. The following points are important:

(1) There is a distinction between *statistical significance* and *clinical significance*. A result may not have arisen by chance; that is, it may reflect a true difference, but be so small as to render it of no practical importance.
(2) It is best to report the actual probability of obtaining the result by chance alone under the null hypothesis, that is, the *actual p value*, rather than just saying it is significant or not. The *p* value for what we commonly call "significance" is arbitrary. By custom, it has been taken to be a *p* value of 0.05 or less. But the 0.05 cutoff point is not sacred. The reader should decide what strength he or she will put in the evidence provided by the study, and the reader must have the information to make that decision. The information must include the design of the study, the sample selection, the sample sizes, the standard deviations, and the actual *p* values.

In summary:

(1) Finding *no significant difference* from a small trial tells us nothing.
(2) Finding *no significant difference* in a large trial is a real finding and tells us the treatments are likely to be equivalent.

(3) Finding a *significant difference* in a small trial may or may not be replicable.
(4) Finding a *significant difference* in a large trial is to be trusted as revealing a true difference.

6.9 What Is Involved in Sample Size Calculation?

(a) Effect Size

Let us say that 15% of victims of a certain type of heart attack die if they are given drug A and 16% die if they are given drug B. Does this 1% difference mean drug A is better? Most people would say this is too small a difference, even if it doesn't arise by chance, to have any clinical importance. Suppose the difference between the two drugs is 5%. Would we now say drug A is better? That would depend on how large a difference we thought was important. The size of the difference we want to detect is called the *effect size*.

To calculate sample size, you need to know the minimum size of the difference between two treatments that you would be willing to *miss* detecting. Suppose, for example, that in your control group, 30% of the patients without the treatment recover. It is your belief that with treatment in the experimental group, 40% will recover. You think this difference in recovery rate is clinically important, and you want to be sure that you can detect a difference at least as large as the difference between 30% and 40%. This means that if the treatment group recovery rate were 35%, you would be willing to miss finding that small an effect. However, if the treatment rate was 40% or more, you would want to be pretty sure to find it. How sure would you want to be? The issue of "how sure" has to do with the "power" of the statistical test.

(b) Power

Statistical power means the *probability* of finding a real effect (of the size that you think is clinically important). The relationships among power, significance level, and type I and type II error are summarized below:

- *Significance level = probability of a type I error* = probability of finding an effect when there really isn't one. This is also known as alpha or α.
- *Probability of type II error* = probability of failing to find an effect when there really is one. This is also known as beta or β.
- *Power* = 1—probability of type II error = probability of finding an effect when there really is one. This is also known as 1—*beta.*

(c) Sample Size

To calculate sample size, you have to specify your choice of effect size, significance level, and desired power. If you choose a significance level of 0.05 and a power of 0.80, then your type II error probability is 1—power or 0.20. This means that you consider a type I error to be four times more serious than a type II error

(0.20/0.05 = 4) or that you are four times as afraid of finding something that isn't there as of failing to find something that is. When you calculate sample size, there is always a trade-off. If you want to decrease the probability of making a type I error, then for a given sample size and effect size, you will increase the probability of making a type II error. You can keep both types of error low by increasing your sample size. The top part of the table on the next page shows the sample sizes necessary to compare two groups with a test between two proportions under different assumptions.

The second row of the table shows that if you want to be able to detect a difference in response rate from 30% in the control group to 50% or more in the treatment group with a probability (power) of 0.80, you would need 73 people in each of the two groups. If, however, you want to be fairly sure that you find a difference as small as the one between 30% and 40%, then you must have 280 people in each group.

If you want to be more sure of finding the difference, say 90% sure instead of 80% sure, then you will need 388 people in each group (rather than the 280 for 0.80 power). If you want to have a more stringent significance level of 0.01, you will need 118 people in each group (compared with the 73 needed for the 0.05 significance level) to be able to detect the difference between 30% and 50%; you will need 455 people (compared with 280 for the 0.05 level) to detect a difference from 30% to 40% response rate.

The bottom part of the table on the next page shows the impact on sample size of a one-tailed test of significance versus a two-tailed test. Recall that a two-tailed test postulates that the response rate in the treatment group can be *either larger or smaller* than the response rate in the control group, whereas a one-tailed test specifies the direction of the hypothesized difference. *A two-tailed test requires a larger sample size*, but that is the one most commonly used.

(d) Some Additional Considerations

For a fixed sample size and a given effect size or difference you want to detect, maximum power occurs when the event rate is about 50%. So to maximize power, it may sometimes be wise to select a group for study that is likely to have the events of interest. For example, if you want to study the effects of a beta-blocker drug on preventing heart attacks, you could get "more power for the money" by studying persons who have already had one heart attack rather than healthy persons, since the former are more likely to have another event (heart attack). Of course, you might then be looking at a different question, the effect of beta-blockers on survivors of heart attack (which would be a secondary prevention trail), rather than the effect of beta-blockers in preventing the first heart attack (a primary prevention trial). Sometimes, a primary prevention trial gives a different answer than a secondary prevention trial. You may be able to intervene to prevent disease among people not yet suffering from the disease, but your intervention may have little effect on someone who has already developed the disease. Clearly, judgment is required.

Sample size examples				
	Assume	Effect size	Power	Sample size
	Control group response rate =	Detect increase in treatment group at least to	With probability of	*n* Needed in each group
Significance level (one-tailed)				
0.05	30% 30%	40% 50%	0.80 0.80	280 73
	30% 30%	40% 50%	0.90 0.90	388 101
0.01	30% 30%	40% 50%	0.80 0.80	455 118
	30% 30%	40% 40%	0.90 0.90	590 153
Significance level = 0.05, sample sizes for one-tailed versus two-tailed tests				
One-Tailed Two-Tailed	30% 30%	40% 40%	0.80 0.80	280 356
One-Tailed Two-Tailed	30% 30%	50% 50%	0.80 0.80	73 92

6.10 How to Calculate Sample Size for the Difference Between Two Proportions

You need to specify what you think the proportion of events is likely to be in each of the two groups being compared. An event may be a response, a death, or a recovery—but it must be a dichotomous variable. Your specification of the event rates in the two groups reflects the size of the difference you would like to be able to detect.

Specify:

$$p_1 = \text{rate in group1}; \quad q_1 = 1 - p_1; \text{alpha} = \text{significance level}$$
$$p_2 = \text{rate in group 2}; \quad q_2 = 1 - p_2; \text{power}$$
$$n = \frac{(p_1 q_1) + (p_2 q_2)}{(p_2 - p_1)^2} \times f(alpha, \ power).$$

The values of f (alpha, power) for a two-tailed test can be obtained from the table below.

		Values of *f* (alpha, power)			
		.95	.90	.80	.50
Alpha significance level .01	.10	10.8	8.6	6.2	2.7
	.05	13.0	10.5	7.9	3.8
	.01	17.8	14.9	11.7	6.6

Note: n is roughly inversely proportional to $(p_2–p_1)^2$.

Example
Suppose you want to find the sample size to detect a difference from 30% to 40% between two groups, with a power of 0.80 and a significance level of 0.05. Then,

$$p_1 = .30; \quad q_1 = .70; \quad \text{alpha} = .05$$
$$p_2 = .40; \quad q_2 = .60; \quad \text{power} = .80$$
$$f(\text{alpha, power}) = 7.9 \text{ from the table}$$
$$n = \frac{(.30)(.70) + (.40)(.60)}{(.40 - .30)^2} \times 7.9 = 356.$$

You would need 356 people in each group to be 80% sure you can detect a difference from 30% to 40% at the 0.05 level.

6.11 How to Calculate Sample Size for Testing the Difference Between Two Means

The formula to calculate sample size for a test of the difference between two means, assuming there is to be an equal number in each group, is

$$n = \frac{k \times 2\sigma^2}{(MD)^2} = \text{number in each group}$$

where σ^2 is the error variance, *MD* is the minimum difference one wishes to detect, and k depends on the significance level and power desired. Selected values of k are shown on the next page. For example, to detect a difference in mean IQ of 5 points between two groups of people, where the variance $= 16^2 = 256$, at a significance level of 0.05 and with power of 0.80, we would need 161 in each group, or a total sample size of 322. This means we are 80% likely to detect a difference as large or larger than 5 points. For a 10-point difference, we would need 54 people in each group.

$$n = \frac{7.849 \times 2(256)}{(5)^2} = 161 \text{ people}$$

Significance level	Power	k
0.05	0.99	18.372
	0.95	12.995
	0.90	10.507
	0.80	7.849
0.01	0.99	24.031
	0.95	17.814
	0.90	14.879
	0.80	11.679

A common set of parameters for such sample size calculations are $\alpha = 0.05$ and power $= 0.80$. However, when there are multiple comparisons, we have to set α at lower levels as described in Section 3.24 on the Bonferroni procedure. Then, our sample size would need to be greater.

If we are hoping to show that two treatments are equivalent, we have to set the minimum difference we want to detect to be very small and the power to be very, very high, resulting in very large sample sizes.

To calculate values of k that are not tabulated here, the reader is referred to the book *Methods in Observational Epidemiology* by Kelsey, Whittmore, Evans, and Thompson for an excellent explanation. There are computer programs available to calculate power for many different situations. An excellent one is National Council for Social Studies (NCSS) statistical software, which can be obtained by going to the website www.ncss.com.

Chapter 7
Mostly About Quality of Life

The life which is unexamined is not worth living.
Plato, Dialogues (428–348 B.C.)

I love long life better than figs.
Shakespeare (Anthony and Cleopatra)

The two quotes above illustrate how differently people view the quality of their lives and how difficult it is to pin down this concept.

A welcome development in health-care research is the increasing attention being paid to the quality of life issues in epidemiologic studies and when evaluating competing therapies. A key aspect is the measurement of the effects of symptoms of illness, as well as of the *treatment* of these symptoms, on well-being, which is a subjective and relative state. Therefore, it is quite appropriate that measurement of improvement or deterioration in quality of life be based on the *patient's perception and self-report*. A person who has had severe and disabling angina may perceive improved well-being as a result of treatment if he can walk without pain, whereas a young ski enthusiast may experience marked deterioration if he is unable to ski. For that reason, in studies on this issue, the individual often serves as his or her own control, and the measures used are *change scores* in some quality of life dimensions from before to after treatment. However, it remains important to have an appropriate control group to compare the changes, because people show changes in these dimensions over time that may be unrelated to the particular treatment being evaluated.

The principles and techniques described in this book apply to research in any health-related field. However, there are certain analytic methods that are particularly appropriate to investigations concerning psychological or emotional states. The primary principle is that if it is to be scientific research, it must adhere to scientific standards, which means that first of all, *the variables of interest must be quantified*. Fortunately, almost any concept related to the health fields can be quantified if one is ingenious enough.

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_7

7.1 Scale Construction

The scales used to measure quality of life dimensions reflect the degree of distress with particular symptoms or psychological states as well as degree of satisfaction and general well-being. There are many such scales available, which have been well constructed and tested on different populations. Sometimes, however, investigators find it necessary to construct their own scales to fit particular circumstances.

There are three characteristics of such scales that are important: *reliability*, *validity*, and *responsiveness*.

7.2 Reliability

Reliability is the ability to measure something the same way twice. It rests on the assumption that a person's score on a scale or a test is composed of his true (but unknown) score plus some component that is subject to variation because of error (by which we mean random variability).

Reliability of a scale is related to its *repeatability*, or how close the responses are on two administrations of the scale. To measure how close they are, we can calculate the correlation coefficient between the two administrations of the scale to the same subjects. But often we can't give the same scale to our patients twice under exactly the same circumstances, since in reality a patient responding twice to the same questions would respond differently either because something has intervened between the two occasions or because he remembered the previous responses or just because there is inherent variability in how one feels. The next best thing would be to give two equivalent scales to the same group, but that has its problems as well. How do we know the two scales are really equivalent?

Fortunately, there are various measures of what we call "internal consistency" that give us the reliability of a scale or test. The most common one is called Cronbach's alpha. There are many software packages for personal computers that readily calculate Cronbach's alpha, including SPSS, SAS, STATA, and many others. Thus, it is not necessary to calculate it yourself, but the following explanation indicates what it really means and how to interpret it.

$$\alpha = \left[\frac{k}{k-1}\right] \times \left[\frac{variance\ \ of\ \ total\ \ scale - sum\ \ of\ \ variances\ \ of\ \ individual\ \ items}{variance\ \ of\ \ total\ \ scale}\right].$$

Variance is the standard deviation squared. Section 3.4 shows how to calculate it. (When we talk about variance here, we actually mean the population variance, but what we really use are estimates of the population variance that we get from the particular sample of people on whom we develop the test or scale, since obviously we can't measure the entire population.)

This formula is really a measure of how homogeneous the scale items are, that is, to what extent they measure the same thing. If you have a scale that is composed of several different subscales, each measuring different things, then the Cronbach's alpha should be used for each of the subscales separately rather than the whole scale. Cronbach's alpha gives the lower bound for reliability. If it is high for the whole scale, then you know the scale is reliable (repeatable, highly correlated with the "true," but unknown, scores). If you get a low alpha for the whole scale, then either it is unreliable or it measures several different things.

Reliability can also be looked upon as a measure of correlation, and in fact it does reflect the average correlation among items of a scale, taking into account the total number of items. Another way to get reliability is from the Spearman–Brown formula, which is

$$\frac{k(average\ \ correlation\ \ among\ \ all\ \ items)}{1+(k-1)\ \ average\ \ correlation\ \ among\ \ all\ \ items} = \frac{k\left(r_{average}\right)}{1+(k-1)\left(r_{average}\right)}.$$

As this formula indicates, a longer test or scale is generally more reliable if the additional items measure the same thing. On the other hand, shorter scales are more acceptable to patients. An alpha above 0.80 is considered very good, and sometimes subscales are acceptable with alpha over 0.50, particularly when there are a large number of subjects (over 300), but it should be considered in the context of the other psychometric qualities of the scale.

There are other measures of reliability as well. Psychometrics is a specialized and complex field and there are many excellent books on the subject, for example, *Health Measurement Scales* by Streiner and Norman.

7.3 Validity

Validity refers to the degree to which the test measures what it is supposed to measure. An ideal situation would be one in which there is some external criterion against which to judge the measuring instrument, a "gold standard." For example, if it could be shown that anxiety as measured on one scale correlates better with some objectively definable and agreed-upon outcome than anxiety measured on a second scale, one could say the first scale is more valid. (This is called "criterion validity.")

Unfortunately, in quality of life issues, there are generally no external criteria. A person may feel he or she is miserable, but may be functioning at a high level. The very idea of quality of life is conceptually subjective. Whose quality of life is it anyway?

Therefore, we often must rely on content validity, which is a blend of common sense and technical psychometric properties. If we want to know if someone feels depressed, we might ask, "Do you feel sad a great deal?" rather than, "Do you feel athletic?" However, even that is not so simple, since what someone who is not an expert on depression might consider overtly irrelevant, like sleep disturbances, is one of the most powerful signs of depression.

Of course, if there is an external criterion against which to validate a scale, it should be used. But even content validity may be made more objective, for instance, by forming a group of experts to make judgments on the content validity of items. To test the agreement between judges, the kappa coefficient may be used, as described in Section 3.3.

7.4 Responsiveness

Responsiveness of a scale is a measure of how well it can detect changes in response to some intervention. Responsiveness, or sensitivity of a scale, can be assessed in several different ways and there is no consensus as to which is the best. Some related concepts are described below, which pertain to the situation when you are looking at change from pre- and posttreatment measures.

(1) The use of *change scores* (pre-post) is appropriate when the variability between patients is greater than the variability within patients. In general, change scores can safely be used if

$$\frac{\sigma^2_{between\ \ patients}}{\sigma^2_{between} + \sigma^2_{error}} \geq 0.5$$

$\sigma^2_{between\ patients}$ and σ^2_{error} can be obtained from an analysis of variance of scores of a group of patients who have replicated measures, so that you can estimate the variance due to error.

(2) A *coefficient of sensitivity* to change due to a treatment is

$$\frac{\sigma^2_{change}}{\sigma^2_{change} + \sigma^2_{error}}.$$

To get the σ^2_{error}, one needs to do an analysis of variance of repeated measures on the same subjects. Computer programs are available. Detailed explanations of this appear in more advanced texts.

(3) *Effect size* is simply the change in the scale from before to after treatment, divided by the standard deviation at baseline. The standard deviation is an index of the general variability in scores among the group of people in the study. One can measure the magnitude of the average change in scores after some treatment by determining what percentage of the "background variation" that change represents. *Effect size* =

$$\frac{mean\ \ change\ \ score}{standard\ \ deviation\ of\ \ baseline\ \ (or\ \ pretreatment)\ \ scores}.$$

(4) A *measure of responsiveness* proposed by Guyatt et al.[30] is

$$\frac{mean\ change\ score}{standard\ deviation\ of\ \ change\ scores\ for\ ''stable\ subjects''}.$$

"Stable subjects" are hard to define, but what this suggests is that a control group that doesn't get the intervention or gets placebo may be used. Then one can use the standard deviation of the change scores in the control group as the denominator in the term above.

The variability of the change scores in the control group (or in a group of stable subjects) can be looked at as the "background variability" of *changes*, and the measuring instrument is responsive to the degree it can detect changes above and beyond this background variability.

(5) When evaluating change due to treatment, one should *always* have a control group (i.e., a no-intervention or placebo group) for comparison, since change can occur in control patients as well, and the question of interest is whether the pre- to posttreatment change in the treatment group exceeds the "background" change in the control group. If you use effect size as a measure, then you should compare effect size in the treatment group with effect size in the control group.

A numerical example of these concepts is provided in Appendix 6.

7.5 Some Potential Pitfalls

(a) Multiplicity of Variables

Quality of life research often deals with a vast quantity of variables. Let us say an investigator is examining the effects of a drug to treat hypertension and comparing it with placebo. The investigator may have several hundred items to assess various physical and psychological symptoms and side effects. If one were to compare the two groups by *t*-test on each of the items, at the $p = 0.05$ level of significance, one would expect that roughly 5% of these tests would produce a significant result by

chance alone. The exact probability is difficult to determine, since some of these comparisons would be correlated by virtue of the fact that the same patients are responding to all of them; that is, the responses are not independent. But in any case, if the investigators pick out just the significant items and conclude that there are effects of the drug, they may be committing type I errors, that is, rejecting the null hypothesis incorrectly.

That is why it is important to use scales that measure particular constructs or to group items in a clinically meaningful way. For example, one might wish to measure depression, anxiety, hostility, and well-being (each of which consists of multiple items). On the other hand, certain drugs may be related to very specific symptoms, such as nightmares, and this might need to be assessed by a single item that asks about the frequency of nightmares.

The point is that quality of life research should generally be driven by some specific hypotheses. Otherwise, it becomes a "fishing expedition" that just fishes around for anything significant it can find. It should be noted that "fishing expeditions" may be useful to generate hypotheses that then need to be tested in a different study.

(b) Generalization

Another important issue is the extrapolation of results to populations other than the one from which the study sample was drawn. Quality of life effects may be different in men than in women, in younger than in older people, and may differ by ethnic and cultural groups. One should be careful in making generalizations. In addition, psychometric properties of a scale in one language may not be the same as in another language, so the researcher must be careful to ensure that a translated scale has the same meaning as the original scale. A psychometrician should be consulted when constructing a scale.

(c) Need for Rigorous Standards of Research

Some people consider quality of life measures "soft." What they generally mean is that they think such measures are subjective, variable, and perhaps meaningless. That is nonsense, and to the extent it is true in some studies, it reflects the inadequacies of the researcher, not of the subject matter. These measures *should be subjective* from the patient's perspective, since they reflect the patient's subjective perception of well-being or distress. *It is the researcher who should not be subjective* and who need not be if he follows the principles of research. The variability in quality of life measures is no greater than in many physiological measures and, in any case, is part of the essence of some quality of life constructs. As for meaning, that is a philosophical issue, not a scientific one. From the scientific viewpoint, the "meaning" should be defined operationally. Quality of life research should adhere to the principles of all good research and the general approach is the same as for any scientific investigation:

(1) Formulate a testable hypothesis.
(2) Quantify the dependent variable (or variables).

(3) Select a study design that can answer the question you've posed.
(4) Quantify the independent variables.
(5) Control for potential confounders (through study design and/or data analysis).
(6) Plan for a sample size that will give you enough power to detect an effect size of interest.
(7) Try to ensure that you minimize systematic bias.
(8) Collect the data, paying much attention to quality control.
(9) Analyze the data using appropriate statistical techniques.
(10) Make inferences consistent with the strengths and limitations of the study.

Chapter 8
Mostly About Genetic Epidemiology

Let us then suppose the mind to be, as we say, white paper (tabula rasa), void of all characters without any ideas; how comes it to be furnished? Whence comes it by that vast store, which the busy and boundless fancy of man has painted on it with an almost endless variety?. . . .To this I answer, in one word, From experience: in that all our knowledge is founded. . . .

John Locke
An Essay Concerning Human Understanding

8.1 A New Scientific Era

We are a long way from believing that the mind is a "tabula rasa," a blank slate. We know now that much is in fact innate, i.e., under genetic influence. The purpose of this chapter is to help those who wish to read the rapidly expanding literature in genetic epidemiology. Thus, it is an overview of the basic designs and statistics used in this area; it is not comprehensive, nor is it highly technical.

The focus of epidemiologic research has evolved as parallel progress has been made in other fields of medicine and basic science. In the era when infectious diseases were rampant, epidemiology was concerned with identifying the sources of the infection and methods of transmission, largely through fieldwork. As the infectious agents were discovered, as sanitation and health status improved, chronic diseases, such as heart disease and cancer, became the leading causes of death and disability in the developed world and came to be the foremost targets of epidemiologic research. (Now that new infectious diseases are once again emerging, this part of epidemiology is again gaining prominence.)

The objective of chronic disease epidemiology was to identify risk factors for these diseases. This part of the story has been a great public health success. We now know, because of epidemiologic studies, what the major modifiable risk factors are for cardiovascular disease: hypertension, high cholesterol and LDL-C (low density lipoprotein cholesterol), smoking, overweight, and inactivity. Our challenge now is to find ways to make the lifestyle changes in the population, which will further lower

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_8

the rates of cardiovascular disease. We also know many of the exposures related to cancer, but not as comprehensively as for heart disease.

At this scientifically historic time, as science is fully entering into the era of *genomics*, *epigenomics*, and *proteomics* (and other "omics"), epidemiology has entered a new phase of research activity: *molecular epidemiology*. This is the search for blood or tissue biomarkers and genetic polymorphisms (variants) that are associated with or predispose to disease. Why is this different from any other risk factor investigated in epidemiology? In many ways it isn't, especially with regard to the blood biomarkers, but in genetic epidemiology, there are study designs and statistical analysis methods that are quite different. A welcome aspect of molecular and genetic epidemiology is the true collaboration of basic scientists, clinicians, and epidemiologists. For too long the disciplines have gone their separate research ways and scientists read mostly the scientific journals in their own field. But molecular epidemiology cannot fruitfully proceed without the interface of laboratory scientists and population researchers.

Below are some basics of genetics that you can skip reading if this is all familiar. Deoxyribonucleic acid (DNA) is made up of four units—or nucleotides. These nucleotides, also called bases, are adenine, guanine, thymine, and cytosine and are denoted by the letters A, G, T, and C. The DNA is arranged in two strands twisted in a double helix form, such that the nucleotides AGCT pair with each other in fixed ways. An A always pairs with T and C always pairs with G. These are called base pairs. If one strand of the double helix were strung out in a line, it might look like this:

AATTCGTCAGTCCC.
TTAAGCAGTCAGGG.
The other strand that pairs with it would be (Figure 8.1).

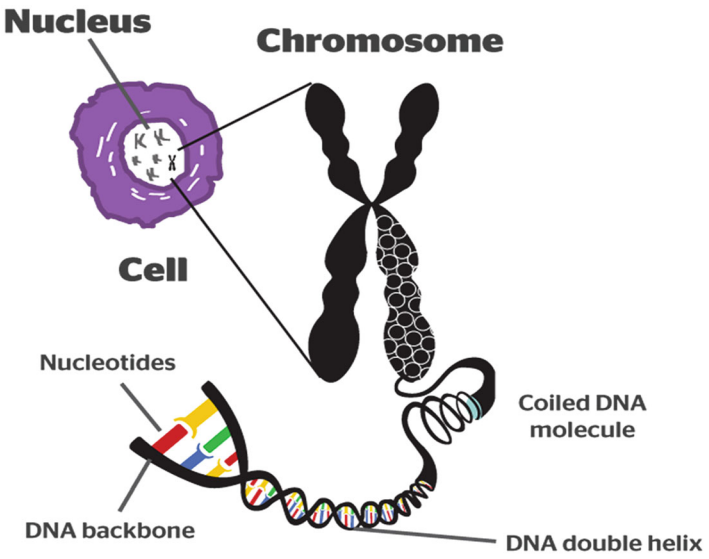


Figure 8.1 DNA

There are three billion base pairs (or six billion bases) in the human genome (which refers to all the genetic material in humans). These three billion base pairs are organized into 23 chromosome pairs (1 from the mother and 1 from the father), which are in every living cell in the body (except the sperm and egg cells, each of which have 1 chromosome each until they merge to form a fertilized cell that now has the full complement of chromosomes). Within these three billion base pairs, there are about 20,000 protein-coding genes that are sequences of base pairs of different lengths and that provide the code for the formation of proteins. A much larger number of genes and sequences serve mostly regulatory functions or have functions that are unknown at the present time. The most common variations in the genome are known as single nucleotide polymorphisms (SNPs, pronounced as "snips") and involve a variation in a single letter of the genetic code. Some SNPs protect against disease, while others may predispose to disease.

Genes are named by HUGO, the international Human Genome Organization. The HUGO Gene Nomenclature Committee (HGNC) provides guidelines for naming genes and their symbols. Figure 8.2 illustrates the meaning of the name: 7q31.2. It is the cytogenetic location on the chromosome of the CFTR gene, which is the gene mutated in cystic fibrosis. The symbols 7q31.2 mean that it is on the seventh chromosome, in the long arm of the chromosome (q), in band 1, sub-band 2.

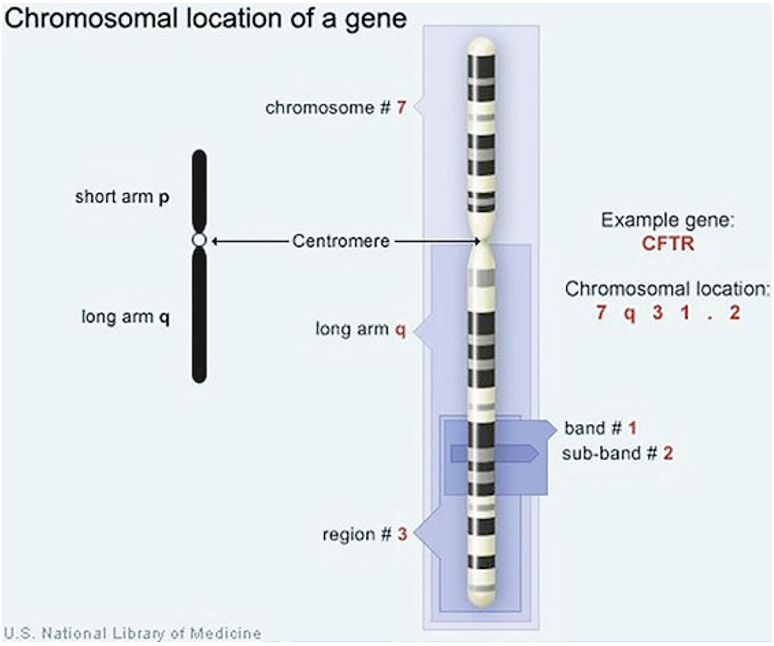


Figure 8.2 Chromosomes and genes.
http://ghr.nlm.nih.gov/handbook/howgeneswork/genelocation

SNP names are assigned by the National Center for Biotechnology Information database of Single Nucleotide Polymorphisms (NCBI dbSNP) group. For example, a SNP name might be rs2532087. The "rs" stands for "reference SNP" and refers to the reference sequence, produced by the Human Genome Project (HGP), that is a composite of data from a number of anonymous volunteers sequenced by the HGP. The number, 2532087, is assigned by NCBI dbSNP group. These numbers are simply identifiers (not genomic positions).

8.2 Overview of Genetic Epidemiology

Genetic epidemiology seeks to identify genes related to disease and to assess the impact of genetic factors on population health and disease. Here is an overview of the strategy often used to study genetic determinants of disease. First we may want to determine if a trait or disease *runs in families*. If it is not familial, it is not likely to be heritable; if it is familial, it may or may not be due to genetic factors (environments run in families also). Next, we want to see if genetic variation contributes to the familial transmission. One method for determining this is by studying twins (described in Section 8.3). If we determine the *disease is heritable*, we would want to *localize and identify* the genes involved. And finally, we want to understand the mechanisms by which these genes contribute to disease.

As a first step, we may want to find out where the genes that contribute to the disorder are located. One approach to this is to conduct linkage studies of individuals affected with the disease and their families (described in Section 8.4). *Linkage studies* may identify regions on the chromosome that are likely to harbor the disease genes. Once we've identified one or more such regions, we may look to see what genes are known to reside in those regions. We can then test these genes using *association studies in families or in unrelated individuals* (described in Section 8.6) to determine whether any variants (also called *alleles*) of these genes are associated with the disease. Another approach, better suited to common complex diseases, is to begin with genome-wide association studies (discussed in Section 8.9).

So, there are a variety of designs and statistical tests that can be used to define the genetic basis of a disease, including (1) *twin studies* to determine if the disease has a heritable component; (2) *linkage studies* to identify and locate regions of chromosomes containing genes involved in the disease; and (3) *association studies* to determine whether specific genetic variants are associated with the disease, to examine how they interact with the environment, and to determine how they affect population health. We will limit the discussion to some pretty simple models that will give the flavor of the topic.

8.3 Twin Studies and Heritability

To explore a genetic influence on disease, we may first look to see if it runs in families. But something that is familial is not necessarily heritable. For example, are obese parents more likely to have obese children because of genetics or because of nutrition and activity levels that are transmitted from the parents to the children? What we want to know is whether and to what extent the phenotype (what we observe in the person, e.g., obesity) is affected by genetic factors.

One way to assess the influence of genetic variation is by studying twins. Identical twins (monozygotic—coming from the same fertilized egg) share 100% of their genes, while fraternal twins (dizygotic—coming from two fertilized eggs) share on average 50%, just as non-twin siblings do. One way to estimate the strength of genetic influences is to calculate the *heritability*, h^2. For twin studies, heritability can be calculated as twice the difference between the correlation for that trait among monozygotic twins minus the correlation in dizygotic twins or

$$h^2 = 2(r_{mz} - r_{dz})$$

Consider blood pressure. If variation in the condition or trait under investigation were completely attributable to genetic variation, then each member of a monozygotic twin pair would be equally affected (each member would have the same blood pressure) and the correlation between monozygotic twins would be 1.0; the correlation in dizygotic twins, however, would be 0.50.

In this case, h^2 would be 2(1–0.5) = 1.0 or 100%. If genetic variation had no effect, then we would expect blood pressure correlations for monozygotic and dizygotic twin pairs to be the same (i.e., $r_{mz} = r_{dz}$) and h^2 would be zero. In reality, diseases and traits are generally partially heritable—that is, h^2 lies somewhere between 0 and 1.0.

If we are talking about continuous variables, we can think of heritability in terms of correlation coefficients. If we are talking about categorical variables, we may speak of concordance rates, where

$$h^2 = \frac{\%\ monozygotic\ twins\ concordant\ for\ the\ disease - \%\ of\ dizygotic\ twins\ concordant}{1 - \%\ of\ dizygotic\ twins\ concordant}$$

Some reported approximate estimates[31–32] of heritability from twin studies are 0.60 for alcoholism, 0.30–0.50 for personality traits, 0.35 for colorectal cancer, 0.26 for multiple sclerosis, 0.75 for height, and 0.80 for schizophrenia.

It is important to remember that heritability doesn't measure how much of an individual's disease is attributable to genetics; rather it tells us what proportion of the population's variability in the phenotype is the result of variation in the genes in the population. So it is a measure applicable to a population, not to an individual. If you have people living in exactly the same environment, then any variation you encounter in the phenotype would be mainly due to genetic factors, and heritability would

approach 100%. So there are some limitations to this measure, but it does give us an idea to what extent genetic variation contributes to phenotypic variation in a population. However, heritability tells us nothing about what genes are responsible for that variation, which genetic variants are involved, how many variants are involved, or what their effect sizes are. This more detailed information is referred to as the "genetic architecture" of a trait or disease.

In recent years, heritability is more commonly estimated from genome-wide SNP data that can capture the heritable component of traits attributable to common variation (SNPs) (see Section 8.16).

8.4 Linkage and Association Studies

If we know that a disease is heritable, we can now turn to the task of actually identifying the genes that are involved. Most disorders that are studied by epidemiologists (e.g., cardiovascular diseases, psychiatric disorders, common forms of cancer) are considered "complex" disorders. Unlike single-gene or Mendelian disorders, such as cystic fibrosis or Huntington's disease, complex disorders are thought to result from the contribution of several or many genes interacting with environmental risk factors. That can make identifying the effect of an individual gene quite a difficult task. The effect of a particular allele within that gene may be quite small. It is a bit like looking for the proverbial needle in the haystack. Nevertheless, genes contributing to diseases are being discovered and there are certain strategies that are employed in the search.

Where in the genome do we look for the genes that confer susceptibility to the disease? One way to answer this question is to use *genetic linkage analysis*.

(a) Linkage analysis relies on the phenomena of crossing over and recombination that occur during the process of meiosis, a special kind of cell division that occurs when the sex cells (sperm and egg) are formed. Each person has two copies of each of the 23 chromosomes that make up the genome: One copy is inherited from the mother and one from the father. During the formation of sperm and egg cells, these 23 chromosome pairs line up and exchange segments of genetic material in a process known as *crossing over*. This recombination occurs at one or more places along the chromosome. *The closer two loci are on a chromosome, the less likely a recombination event will occur between them and so the more likely they will be inherited together*. Loci that tend to be co-inherited are said to be genetically linked. We can use this fact to estimate the distance between two genetic loci or markers (a genetic marker is a DNA variation whose chromosomal location is known). The physical distance between two markers is inversely related to how frequently they are co-inherited across generations in a family.
(b) The *distance between two loci is sometimes measured in centimorgans*. A centimorgan (cM) is a unit of distance along a chromosome, but not in the ordinary sense of physical distance. It is really a probability measure, which is a

reflection of the physical distance; it represents the probability of two markers or loci being separated (or segregated) by crossing over during meiosis. If the two markers are very close together, they are unlikely to separate (we say they are "linked"); if they are far apart, they are likelier to cross over and *the genetic material gets recombined during meiosis*. Then this recombined DNA gets transmitted to the offspring. Two loci are one centimorgan apart if the probability that they are separated by crossing over is only 1% (once in a hundred meioses). It has been estimated that there are about one million base pairs in a 1 cM span. Loci that are far apart, say 50 cM, will be inherited independently of each other, as they would be if they were on different chromosomes. *The purpose of linkage studies is to localize the disease susceptibility gene to be within some region on the chromosome*.

(c) So we might begin our search for a disease gene by collecting families affected by the disease and performing a linkage analysis using markers spaced at intervals (say 10 cM apart) across the entire genome. If we find a marker that appears to be co-transmitted with the disease, we would have evidence that the marker is genetically linked to a gene for the disease. In other words, there is likely to be a gene for the disease in the same region as the linked marker.

(d) Having found a chromosomal region linked to the disease, we might try to narrow the region down by genotyping and testing additional markers within that region (say at 1 cM intervals). However, even this relatively small region may contain many genes.

(e) Our next step might be to screen the genes that are known to reside in this region. We would be particularly interested in genes that have a plausible connection to the disease of interest (these would be good "candidate genes"). For example, if we are studying diabetes, genes that make proteins involved in glucose metabolism would be important "candidate genes."

(f) Now we can see if any particular alleles (variants) of the genes in that chromosomal region are associated with the disease. This can be done by:

Association studies using case–control methods in unrelated people, examining whether an allele is more common in cases than controls (described in Section 8.6). *Association studies in families* to see whether an allele is being transmitted more commonly to cases than expected by chance. So, essentially, linkage analysis tells us that a particular marker location is near a disease susceptibility gene; association analysis tells us that a particular allele of a gene or marker is more commonly inherited by individuals with the disease.

8.5 LOD Score: Linkage Statistic

The classic statistic used to evaluate the strength of the evidence in favor of linkage of a genetic marker and disease susceptibility gene is the LOD score (the $\log_{10}$ of the odds in favor of linkage). It will be described in principle only, to help in interpretation of linkage studies. The actual calculations are complex and require special software packages.

The principle underlying the LOD score is described in the previous section: if we have two loci—say, a marker and a disease gene—the closer they are on a chromosome, the lower the probability that they will be separated by a recombination event during meiosis and the more likely they will be co-inherited by offspring.

The probability of recombination, called the recombination fraction, is denoted by the symbol θ and depends on the distance between the gene and the marker. *If there is no recombination and the gene and marker are completely linked, then the recombination fraction is 0*. The maximum value of θ is 0.5 (if gene and marker were independently inherited, then the probability that the marker was transmitted but not the gene = the probability that gene was transmitted but not the marker = 0.50).

So if you want to know if there is linkage, we have to estimate how likely it is that θ is less than 0.5, given the data we have observed. We use the likelihood ratio for this, which as you recall from Chapter 2 is the ratio of the probability of observed symptoms, given disease divided by the probability of observed symptoms given no disease. In this case

$$LR = \frac{Probability\ \ observed\ \ inheritance\ \ data,\ \ given\ \ linkage}{Probability\ \ observed\ \ inheritance\ \ data,\ \ given\ \ no\ \ linkage}$$

The null hypothesis here is no linkage (or recombination fraction $\theta = 0.5$) and the alternate hypothesis is linkage (or $\theta < 0.5$). If we reject the null, we "accept" the alternate hypothesis. The test statistics used to see if we have sufficient data to conclude linkage is the *LOD score*, *which is the* $log_{10}(LR)$. For Mendelian (single-gene) disorders, a LOD score of 3 has traditionally been the threshold for declaring significant linkage, although for complex disorders higher thresholds (3.3–3.6) have been recommended. A LOD score of 3.0 indicates 10^3 odds in favor of linkage compared to no linkage, i.e., 1000:1 odds in favor of linkage.

For complex reasons beyond the scope of this book (but described in the references at the end), a LOD score can be translated into probability by multiplying it by the constant 4.6: LOD × 4.6 is distributed as chi-square with one degree of freedom. (The 4.6 is 2 times the natural log of 10.) Thus, a LOD of 3.0 is equivalent to a chi-square of $3 \times 4.6 = 13.82$ and corresponds to $p = 0.0002$. The inheritance data for linkage analyses can come from family pedigree studies, from sibships or other family groups.

LOD score linkage analysis is sometimes referred to as "parametric" linkage analysis because it requires that we specify certain parameters (e.g., disease and marker allele frequencies, recessive vs. dominant mode of inheritance, penetrance of the disease gene). When these parameters are known or can be approximated, parametric LOD score analysis is the most powerful method of linkage analysis. This may be true for Mendelian (single-gene) disease, but for many complex disorders, these parameters are not known. "Nonparametric" linkage methods (known as the allele-sharing approach) are often used to study complex disorders because they do not require knowledge of the mode of inheritance or other genetic parameters. There are a number of statistics available, described in the more advanced texts.

8.6 Association Studies

Compared to linkage analysis, association studies are more closely akin to traditional epidemiologic studies and most often rely on the case–control design. In an association study, investigators are interested in finding whether there is any association between a particular allele at a polymorphic locus and the phenotype in question. (A polymorphism is a variation in DNA sequence that occurs in at least 0.5%–1% of the population. Variations that are more rare are usually referred to as mutations.) For the purposes of this discussion, we will assume that the polymorphisms we are looking at are *single nucleotide polymorphisms (SNPs)* or variants in a single one of the bases A, T, C, and G at a particular locus. Note that there are many other classes of DNA variation including small insertions and deletions of nucleotides and copy number variations that involve deletion or duplication of larger chunks of DNA, as illustrated below (Figure 8.3).

Figure 8.3 Variants in the genetic code. Reprinted by permission from Macmillan Publishers Ltd.: [Nature Reviews Genetics] (Frazer KA, Murray SS, Schork NJ, Topol EJ, Human genetic variation and its contribution to complex traits), copyright (2009)

So let us say at a particular SNP some people have the A allele and other people have the G allele. We want to know if people with the disease are more likely to have, say, the G allele than the A allele. For a binary phenotype (e.g., diseased or not), we can do case–control studies of association by taking cases who are affected with the disease and unrelated controls who are not. Remember that each person gets one copy of an allele at a particular locus from the mother and one copy from the father. (These are exactly at the same locus on each chromosome of the paired chromosomes.) So if a person gets an A from the mother and a G from the father, that person's genotype is AG. If the person gets an A from each parent, that person's genotype is AA. We can compare the frequencies of genotypes AA, AG, or GG between cases and controls, represented by the numbers 0, 1, or 2 for the number of

minor alleles it contains (a major allele is the more common one in the population, a minor allele is the less common one). So in our example, let's say the G allele is the minor allele; then we can convert the genotypes to numbers as follows: 0 for the AA genotype, 1 for the AG genotype (since it contains one G), and 2 for the GG genotype. We can then see whether the number of minor alleles (0, 1, or 2) differs between the cases and controls. We can use ordinary statistical tests of the differences between proportions or multiple logistic regressions (see Section 4.18) to determine the odds ratio connected with the allele in question, and we can test for gene–environment interactions by including an interaction term of the presence of the allele and some environmental factor, such as smoking. For a continuous phenotype (e.g., blood pressure), we can use linear regression.

Association studies can be more powerful than linkage analysis for detecting genes of modest effect, making them an attractive approach for studying complex disorders, which involve many risk variants (likely thousands) of relatively small individual effect. Power calculations for association tests can be conducted using several online tools.

8.7 Candidate Gene Association Studies

Association studies may be used to evaluate a candidate gene—that is, a gene that previous data, often from linkage studies or prior biological studies, have suggested is involved in our phenotype of interest. However, our knowledge of the biological basis of many diseases is incomplete. Thus, the prior probability that any given candidate gene or polymorphism is actually involved is often low. In this circumstance, a significant association is likely to be due to chance. Indeed, candidate gene findings have been notoriously difficult to replicate, suggesting that most reported associations are spurious. As a result, candidate gene studies have largely fallen into disfavor and have been replaced by genome-wide association analyses (GWAS, discussed in Section 8.9) that do not rely on prior hypotheses.

8.8 Population Stratification or Population Structure

One important issue for association studies is the choice of control groups. The frequency of alleles in a population can differ based on ancestry for reasons that are unrelated to the disease of interest. Such differences in genetic background may confound our results due to a phenomenon referred to as "population stratification."[33]

For example, let us say the A variant is more common in individuals of European descent and the G variant is more common among those of African descent. If it happens that our disease (case) group has more European–Americans and our

no-disease (control) group has more African–Americans, then we might find that allele A is more common among those with disease than in those without disease, but it might really just be a reflection of the fact that we had more European–American cases than European–American controls and the European–Americans are more likely to be carriers of the A allele. So we would have a false-positive finding because of the unmatched ancestral similarity of the two groups.

The numerical example below illustrates this point. You can skip reading the example if you like, but it is shown here for those who want to work through it numerically. Let us assume that 80% of people of European ancestry[1] carry allele A at a specific SNP and only 40% of people of African ancestry carry allele A. The respective percentages for allele G are 20% in Europeans and 60% in African ancestry (Table 8.1). Now let us say that it happens that among the cases of disease in our sample, 70% are of European descent, but in the controls, only 10% are of European descent (Table 8.2). What % of A alleles would we expect in the cases and controls?

Table 8.1 Allele frequencies of a hypothetical SNP by population ancestry

	Ancestry group	
	European–American	African–American
Prevalence of A allele	80.0%	40.0%
Prevalence of G allele	20.0%	60.0%

Table 8.2 Hypothetical disease prevalence by population ancestry

Disease	Cases	Controls
% European–American	70.0%	10.0%
% African–American	30.0%	90.0%

Well, among the cases, since 70% are of European ancestry and 80% of those carry the A allele, we would expect $0.70 \times 0.80 = 0.56$ or 56% of the European ancestry cases to carry the A allele; the remaining 30% of the cases are of African ancestry in whom the A allele frequency is 40%, so we would expect another $0.30 \times 0.40 = 0.12$ or 12% to carry the A allele, for a total of 56% + 12% = 68% of cases carrying the A allele, as shown in Table 8.3.

[1] Here we use the terms "ancestry" and "descent" as heuristics. However, the scientifically preferred and more precise terminology would be to refer to quantitative "genetic similarity" between individuals in a GWAS and a specified reference panel (e.g., the 1000 Genomes British individuals) or to say 1000 Genomes GBR-like individuals.

Table 8.3 Calculations

A allele	Cases	Controls
EA	0.70 × 0.80 = 0.56	0.10 × 0.80 = 0.08
AA	0.30 × 0.40 = 0.12	0.90 × 0.40 = 0.36
Total A allele	0.56 + 0.12 = 0.68	0.08 + 0.36 = 0.42
G allele	Cases	Controls
EA	0.70 × 0.20 = 0.14	0.10 × 0.20 = 0.02
AA	0.30 × 0.60 = 0.18	0.90 × 0.60 = 0.54
Total G allele	0.14 + 0.18 = 0.32	0.02 + 0.56 = 0.58
Summary of this table		
	Cases	Controls
Prevalence of A allele	68.0%	42.0%
Prevalence of G allele	32.0%	58.0%

In the controls we would have only 10% of European ancestry of whom 80% carry the A allele, so we would expect 0.10 × 0.80 = 0.08 or 8% of the European ancestry controls to carry A allele; 90% of our controls are of African ancestry, for whom the A allele frequency is 40%, so we expect 0.90 × 0.40 = 0.36 or 36% of them to carry the A allele. Combining the European and African ancestry controls, we expect that 8% + 36% = 42% carry the A allele. Similar calculations are made for the G allele, as shown in summary of Table 8.3.

Now let's assume we have a study of 1000 cases and 1000 controls and there is no real association between the A allele and our disease. Nevertheless, as shown in Table 8.4, the allele frequencies simply based on population differences in our case–control sample would indicate an association (chi-square = 136.6; $p < 0.00001$). So we have to very carefully account for population stratification.

Table 8.4 Example of a case–control study with 1000 cases and 1000 controls

	Cases	Controls	
Prevalence of A allele	680	420	1100
Prevalence of G allele	320	580	900
	1000	1000	2000

There are several options available to deal with the problem of confounding by population stratification. Most commonly, in the era of genome-wide association studies (GWAS), we do this by conducting a principal component analysis that clusters individuals by genetic similarity. We can then select case–control samples of similar ancestry and add our principal components of ancestry as covariates in GWAS regression models. Of note, however, population subgroups may exist even within broad categories like this so that residual confounding may be present. We can quantify the degree of spurious inflation in our genetic association statistics using a genomic control inflation factor, called λ (lambda). We calculate λ as the median of the observed chi-square statistics in our GWAS divided by the expected median of a null chi-square distribution with one degree of freedom (which is 0.456).

In general, a $\lambda < 1.10$ is considered acceptable evidence of no spurious inflation of test statistics. A more recent method called LD score regression (described in Section 8.16) is now widely used to assess confounding due to population stratification. Finally, we can avoid the problem of population stratification by using a family-based[34] rather than case–control design.

If we find an association between a particular allele and the disease we are studying, it may be for one of four reasons: (1) It could be a false positive due to chance (type I error); (2) it could be a false positive due to confounding, for example, because of population stratification; (3) it may be in "linkage disequilibrium" with the true disease allele, meaning that the allele we found more frequently in cases than in controls is located physically close enough to the true causal allele that the two alleles tend to be inherited together and co-occur in affected individuals; and (4) there really is a true causal association of the allele we studied and the disease. (As we said before, genetics—and life—are not simple.)

8.9 Genome-Wide Association Studies (GWAS)

The era of candidate gene study predominance, roughly the decade from 1996 to 2006, was marked by failure: Very few of the thousands of association findings reported in that era were replicable in independent studies. It became clear that a very different approach was needed if efforts to identify genetic variants associated with complex traits were to succeed. Several important lessons were drawn from these failures. First, it became clear that the effect sizes of common variants on complex traits were likely to be vastly smaller than those reported in candidate gene studies to date. As such, the overwhelming majority of reported associations were likely to be false positives (i.e., due to type I error). Second, the biological hypotheses that motivated most candidate gene studies were much weaker than had been presumed. Our knowledge of the biological basis of complex traits is frankly limited, and existing candidate gene studies focused only on a tiny fraction of the millions of variants that were ultimately identified through the sequencing of the human genome. Fortunately, by 2007, technological advances made it feasible to move beyond the candidate gene approach by assaying hundreds of thousands of variants throughout the genome. Rather than investigating a single candidate gene, it was now possible to conduct *genome-wide association studies (GWAS)* to test polymorphisms throughout the whole genome[35]. As a result, GWAS is often referred to as an "unbiased" approach to association—that is, it doesn't rely on prior hypotheses that are usually based on incomplete knowledge of the biological basis of a disease or trait.

A GWAS examines a large set of SNPs (typically a million or more) distributed across the genome to cover common variation (alleles that are carried by 0.5% to 1% or more of the population). These SNPs are typically genotyped using DNA microarrays (sometimes called "SNP chips"), which may also include rarer variants that occur in the exome (the exome consists of the 1–2% of the genome that actually

codes for proteins). More recently, whole genome sequencing, in which all variants (common and rare) are assayed directly, can be used to perform GWAS.

Although the genome contains many millions of SNPs, we can select a smaller subset to cover genome-wide common variation because alleles at many SNPs within a given region are correlated and therefore carry redundant information. This is due to the phenomenon of linkage disequilibrium that we mentioned earlier. Linkage disequilibrium refers to the nonrandom association (i.e., correlation) between alleles that are physically close together on a chromosome and are inherited together. In other words, because of linkage disequilibrium, one SNP may stand in for (or "tag") a larger set of SNPs. As a result, if we find that an SNP is associated with our phenotype, it may or may not be causally related to the phenotype but rather may simply be correlated (i.e., in linkage disequilibrium) with the true causal SNP.

In addition to the SNPs directly genotyped on a microarray, it is now a standard practice to expand the coverage of ungenotyped SNPs by performing imputation. This is made possible by efforts—including the International HapMap Project, the 1000 Genomes Project, and the Haplotype Reference Consortium—that have genotyped or sequenced the genomes of individuals from populations around the world. These samples serve as reference panels that include many more SNPs than those included in a given experiment. Knowing the linkage disequilibrium patterns of SNPs in these reference panels, we can impute alleles at SNPs that we have not been directly genotyped, recovering a much larger number of SNPs for our experiment. GWAS analysis is typically performed using specialized statistical packages, of which the most widely used is known as PLINK (https://www.cog-genomics.org/plink/).

Most of the common SNPs identified by GWAS fall outside of the protein-coding part of genes. Instead, they are found mostly in DNA regions thought to be involved in regulating the activity of genes (gene expression). As noted above, however, an associated SNP may be merely in linkage disequilibrium with the true causal SNP (or it may be a false positive as we discuss later). Identifying the true source of an association signal thus requires additional studies. These can include fine-mapping studies in which the associated region is more deeply genotyped or sequenced (see Sections 8.16 and 8.17) to narrow down the signal, or biological studies in which associated SNPs are examined for their possible functional effect on gene expression, or other biological assays.

8.10 GWAS Quality Control and Hardy–Weinberg Equilibrium

Before imputing genotypes and examining the results of a GWAS, it is crucial to clean the data by performing a series of quality control steps. We typically begin by removing SNPs and individuals for whom data quality is poor. For example, it's customary to *remove SNPs* for which more than say 5% of genotypes are missing

and then *remove individuals* for which more than 2% of genotypes are missing. Next, we may remove SNPs for which the difference between cases and controls in % missing genotypes is greater than 2%; for instance, you would remove SNPs where 10% of genotypes are missing in cases and only 2% are missing in controls. In addition, we remove SNPs that are not in Hardy–Weinberg equilibrium. Hardy–Weinberg equilibrium (HWE) refers to the phenomenon that, under the assumption of random mating, genotype frequencies remain constant across generations, unless some evolutionary disturbances disrupt the equilibrium. This principle is useful for quality control of genotyping, as described below.

Consider one SNP locus with two alleles, say A and G. The frequency of A in the population is p and the frequency of G is q. Since at that locus you can only have either an A or a G, then $p + q = 1$. Now the probability that you get an A allele from your mother and an A allele from your father is the product of the two probabilities since they are independent (see Section 2.2). So the probability that your genotype is AA is

$$p(A_{\text{mother}}) \times p(A_{\text{father}}) = p^2.$$

What's the probability that your genotype has one A and one G, i.e., you are AG or GA? Well there are two ways you can have one A allele and one G allele: you get an A from your mother and a G from your father with probability pq or you can get a G from your mother and an A from your father also with probability of pq. So the probability of having one A and one G is 2pq.

What's the probability of a GG genotype? Here you need to get a G from your father and mother and the probability of a G allele is q so the probability of GG is $q \times q$ or q^2. Thus, there are three possible genotypes, AA, AG, or GG, and the sum of all the possible genotype probabilities at that locus is 1, so $p^2 + 2pq + q^2 = 1$. If you recall your basic algebra, $p^2 + 2pq + q^2$ is the expansion of $(p + q)^2$.

If the genotype proportions in a sample are significantly different from $p^2 + 2pq + q^2$, we need to know why. SNP genotypes may not be in Hardy–Weinberg equilibrium because of genotyping errors or because of violations of the random mating assumption. Thus, we need to remove these SNPs from our dataset. However, in case–control studies, an SNP that is truly associated with the disease may not follow HWE in the combined case–control sample because the alleles have different distributions in cases and controls. Because of the large number of tests performed in a GWAS, the significance thresholds for violations of HWE are typically $p < 10^{-10}$ for cases and $p < 10^{-6}$ for controls. SNPs that violate HWE are removed.

We also want to remove related individuals as they would violate the assumption that our cases and controls are unrelated. We can do this by calculating a statistic called $\hat{\pi}$ (a measure of the proportion of SNPs that are inherited "identical-by-descent" or IBD). For identical twins, $\hat{\pi}$ would be 1.0 (all their alleles are inherited IBD), and for second-degree relatives, $\hat{\pi}$ would on average be 0.25 (which is the proportion of the genome you would inherit from second-degree relatives). In a

GWAS, we typically remove one individual from each pair of individuals for whom $\widehat{\pi}$ is >0.2.

A more extensive tutorial on conducting GWAS is provided in a paper by Marees and colleagues[36] and a bioinfomatic toolkit called Ricopili[37] is widely used to perform quality control, imputation, and GWAS analysis.

8.11 Quantile by Quantile Plots or Q-Q Plots

Constructing a Q-Q plot is often the first step in evaluating the quality of a GWAS. Q-Q plots allow us to visually compare two distributions of test statistics or p values. In GWAS, Q-Q plots are often used to examine if the observed p values follow the distribution expected under the overall null hypothesis. Deviations from the null distribution may indicate problems with the data including inflation of test statistics due to population stratification, problems with imputation, sparse data, or genotyping error.

Suppose you do a GWAS of 100,000 cases and 100,000 controls looking at one million SNPs. Each of those SNP statistical tests between cases and controls has a p value associated with it. The p value comes from a chi-square test comparing the frequency of a given allele in cases vs. controls (or it may come from a logistic regression, where you ascertain the odds ratio of disease [case status] associated with a given allele while controlling for other variables). In any case, you can put those p values for the million SNPs into an ordered list, except that instead of going from the lowest value to the highest value, we reverse it and order the list going from the largest p value to the lowest p value, for example, going from $p = 1.0$ to $p = 0.000001$.

An observed p value of 1.0 is at the 0th percentile—meaning that all p values are less than $p = 1.0$ and there are no values higher than $p = 1.0$ (in our scheme, they are to right of 1.0 in our ordered list; see example below). The p value of 0.000001 may be at the 99th percentile meaning that 99% of the p values are higher than 0.000001 or only 1% of p values are lower than 0.000001 (to the right of 0.000001). In actuality, we transform them to –log p so that p of 1.0 would be 0 and p of 0.000001 would be 6. This is illustrated in the table below.

P	1.00	0.10	0.05	0.01	0.001	0.0001	0.00001
Percentile	0	90th	95th	99th	999th	9999th	99999th
–log10 P	0	1	1.3	2	3	4	5

Now consider the expected distribution of p values if in fact the null hypothesis were true, meaning that there is no significant overall difference between cases and controls, only random variation in the million SNP test p values from chi-square statistics. Now you plot the observed distribution of p value quantiles from your GWAS against the expected theoretical distribution, as below (Figure 8.4).

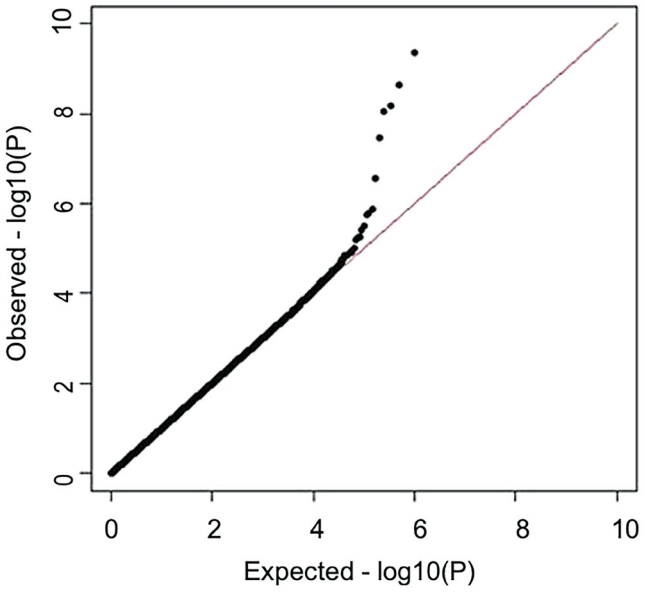


Figure 8.4 Q-Q plot

If the distribution of the p values in your GWAS were exactly like the expected distribution of p values under the null hypothesis, the points would lie along the diagonal red reference line. The points in black are what you actually got from your GWAS. They follow along the red line up until the upper right portion where they are above the red line. This means that you got more points with very, very low p values (or high values of $-\log_{10}$ p) than you would expect to get under the null hypothesis. So there are a bunch of SNP tests that have lower p values than expected and these may represent true SNP differences between cases and controls.

Now look at another Q-Q plot below. Here the points from your GWAS depart from the reference line pretty much along most of the plot, meaning the quantiles from your GWAS do not correspond to the expected quantiles. There are more low p value SNP tests in your GWAS than you would expect along a large portion of the ordered p values. So something is wrong—there are probably too many false positives, maybe due to laboratory errors, population stratification, or other reasons like hidden genetic relatedness among your subjects. This is not a happy situation as it casts doubt on the significant SNP tests you found. What you want to see in a Q-Q plot is points as close to the diagonal line as possible throughout most of the distribution to indicate that there are no errors or biases or confounding between cases and controls. So Q-Q plots are a sort of visual quality control method (Figure 8.5).

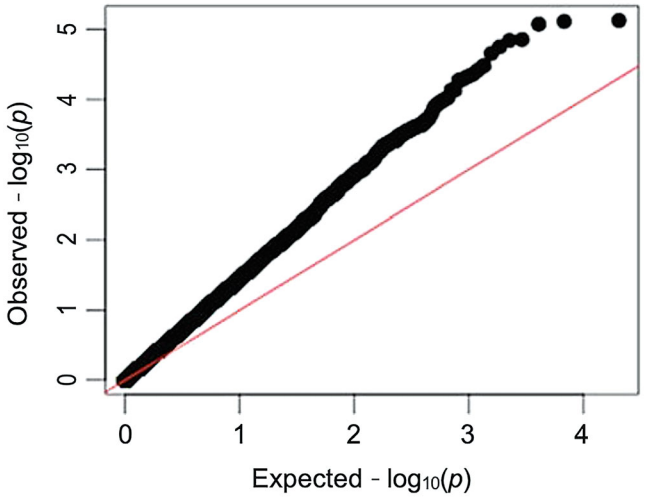


Figure 8.5 Q-Q plot as a means of quality control

We can also quantify inflation of the test statistics using λ (see Section 8.8). In the case of a GWAS study, λ is the median of the chi-squares of tests between cases and controls for all the markers, divided by 0.456, which is the median of the chi-square distribution under the null. A λ of 1.0 indicates no inflation. For the Q-Q plot shown above, λ is 1.1 suggesting minimal inflation. However, for the Q-Q plot shown below, the λ is substantial (1.788) suggesting either many false positives (or possibly that our trait is so highly polygenic and our sample is large enough that there are detectable significant effects across most of the genome).

As we noted earlier (Section 8.8), we can deal with population stratification in GWAS by controlling for genetic ancestry using the many thousands or millions of SNPs that are not associated with the disease you are studying. Essentially, we use these SNPs to capture variation in genetic ancestry and we reduce this genetic background variation to a manageable number of variables (usually 4–20), by applying "principal component analysis" (PCA), anchoring variation in our sample to a previously genotyped or sequenced reference panel of individuals representing diverse populations (such as the 1000 Genomes Project). These principal components of genetic variation can then be added as covariates to logistic regression or linear regression association tests of each SNP. When a GWAS includes the appropriate number of principal components, we are effectively controlling for population stratification, and λ is no longer spuriously inflated.

8.12 Problems of False Positives

Remember that in an association study we are trying to see if a specific allele, say the G allele of an A/G SNP, appears more frequently in cases than in controls. We can use a standard statistical test like chi-square to compare case and control frequencies.

But recall the problem of multiple testing, which leads to a high type I error—i.e., a high probability that we will declare a difference between cases and controls when there really is no difference (Section 3.24). We noted that, for example, if we make five two-way comparisons in a study, the probability that at least one of those comparisons will reach the 0.05 significance level by chance alone is really $1–0.95^5$, or 0.23, which would be our effective significance level. To achieve a true overall significance level of 0.05, we use the Bonferroni correction where we divide 0.05 by the number of comparisons or $0.05/5 = 0.01$, so that we have to reach a level of significance of 0.01 (rather than 0.05) to conclude that there is a significant difference in any of the two-way comparisons.

What correction should we use for a GWAS? It turns out that a typical GWAS study involves effectively one million independent tests. Note that, because of the linkage disequilibrium structure of the genome (i.e., the correlation between nearby SNPs), this effective number of tests is approximately true whether our DNA microarray includes 500,000 SNPs or five million SNPs. If we used the traditional p value threshold of 0.05, without correcting for the multiple tests, we would expect 5% or 50,000 SNPs to exceed that threshold just by chance! So to avoid false positives, we apply the Bonferroni correction and divide 0.05 by one million, which gives us a statistical threshold of 5×10^{-8}. Thus, we need to achieve a p value of <0.00000005 to declare statistical significance of any given SNP association. Of course, if we test multiple phenotypes, we should further correct by the number of phenotypes examined to maintain our type I error rate of 5%. It should also be noted that false positives can occur due to errors in genotyping or even due to insufficient sample size. Thus, genome-wide significant findings are most reliable when they are obtained in well-powered studies after stringent quality control of the dataset. Ultimately, the validity of an association is not established until it has been replicated in independent samples.

8.13 Problem of False Negatives

If you perform a GWAS and find no SNPs that are significantly different in cases versus controls (with $p < 5 \times 10^{-8}$), it may be due to a lack of power. Recall that power is the probability of finding an effect when there really is one and depends on both the effect size of your predictor (here an SNP) and sample size. Typical effect sizes in GWAS give small odds ratios, in the range of 1.01–1.1. The power to detect such an effect depends on several factors: the allele frequency, sample size, disease prevalence (for a binary disease), case/control ratio, and the mode of inheritance of the risk allele (additive, multiplicative, recessive, or dominant). To give an example, detecting an allele with a frequency of 10% that carries an odds ratio of 1.2 for a disease with a population prevalence of 3% under an additive model, we would achieve 90% power with approximately 20,000 cases if we assume an equal number of controls. For an allele with a frequency of 20%, we would need 7000 cases using the same assumptions for effect size, disease prevalence, case/control ratio, and

mode of inheritance. We can also calculate the minimum detectable odds ratio for a given sample size. For example, given 10,000 cases and 10,000 controls and assuming an allele frequency of 10%, a disease prevalence of 3%, and an additive model, we could detect an odds ratio of about 1.23. As these examples make clear, very large sample sizes are needed to detect the kind of effects we expect to see in a GWAS. That is the reason that most genetic studies require collaborations or consortia of several different cohorts or biobanks to achieve sufficient power.

Some have suggested that the threshold for significance of 5×10^{-8} is too stringent and instead of the Bonferroni simple procedure, they use a statistic called false-discovery rate, which is the number of null hypotheses falsely rejected divided by the total number of rejections. There are other nuances in trying to avoid false positives while ensuring the highest probability of finding true positives that are beyond the scope of this book, but references are provided in the reference section.[38]

8.14 Manhattan Plots

A Manhattan plot is a visualization tool that allows us to summarize the entire association results of a GWAS in a single plot. Consider a case–control study of the genetics of ischemic stroke. Each case and each control has been genotyped on a microarray chip containing one million SNPs. How do you make sense of these one million tests? The Manhattan plot is basically a plot of the p values achieved by each SNP you have tested.

On the X axis is the location of the SNPs across the genome, ordered according to the position along the 23 pairs of chromosomes. Imagine the one million SNPs on the chip stretched out in a long line. The Y axis corresponds to the significance level for the association test of each SNP; it is −log of the probability, so that for a p value of 0.0001, −log of 0.0001 = 4; for a p value of 0.00001, −log of 0.00001 = 5; and so on. So the lower the p value, the higher is the number –log(P).

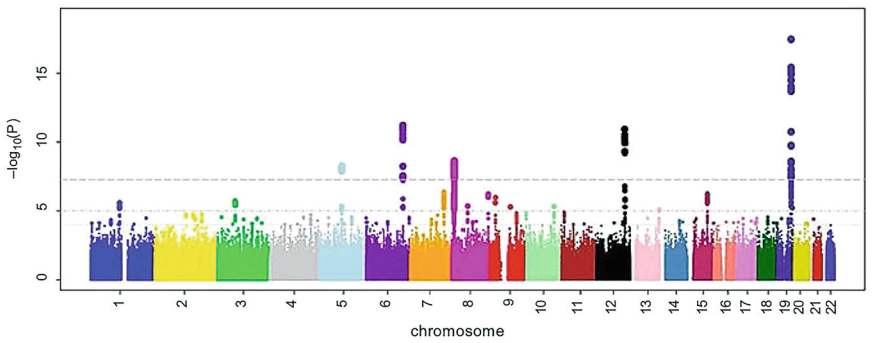


Figure 8.6 Manhattan Plot

The upper horizontal line corresponds to the accepted threshold p value of 5×10^{-8}, which minimizes false positives (see Section 8.12). Anything above that line is considered unlikely enough to have arisen by chance so that we may suppose there is an actual difference between cases and controls at that SNP. In Figure 8.6, we see five genome-wide significant hits—i.e., regions that exceed this threshold (one each on chromosomes 6, 8, 12, and 19). The most significant SNP is at the highest point, and the "tail" of dots directly underneath corresponds to SNPs that are in linkage disequilibrium with the top SNP.

8.15 Polygenic Scores

Complex diseases or traits are now known to be polygenic—that is, many hundreds or thousands of loci may contribute, including both common and rare variants. Thus, we assume that many SNPs that fail to exceed genome-wide significance (i.e., $p < 5 \times 10^{-8}$) are nevertheless truly associated with the phenotype of interest. In addition to discovering new associations by testing each SNP, GWAS can also be used to generate aggregate scores that capture the effects of many alleles in a single variable. These "polygenic scores" can then be used as predictors or covariates in future studies of a disease in independent samples that have been GWAS'd. In other words, the polygenic score provides a single measure of an individual's genetic loading for common risk alleles.

Here's how to construct a *polygenic score*. We begin with a GWAS of the disease of interest, say myocardial infarction (MI), in our primary or "discovery" sample. Then we rank the SNPs by their p values from the lowest to the highest. We might use all SNPs or only use SNPs whose p values are less than a certain threshold (say $p < 0.0001$). Once we have all the SNPs we are going to use, the formula for the score is

$$\text{Polygene score} = \Sigma_{(\text{i to j})}\ x_{\text{i}}\ \log(\text{OR}_{\text{i}})$$

where x is the number of risk alleles (0, 1, or 2) that an individual carries at the ith SNP; OR_{i} is the allelic odds ratio we found for that SNP; and i to j means we sum across all the SNPs included in the score up to the final jth SNP. We also filter SNPs to include those that are largely independent, e.g., by "clumping" SNPs, so that in a given region of LD, we only retain the SNP with the lowest p value. We set a threshold for clumping SNPs based on the degree of LD between them, usually measured by a quantity called r^2 (the squared correlation between alleles at two SNPs). For example, we might remove SNPs whose r^2 is >0.1 with the lead SNP in a region.

We now apply the polygenic score formula to each person in our new, independent "target" dataset to calculate their score across all SNPs. Often, polygenic score weights for a disease have already been calculated from prior GWAS studies and the

weights ($\log(OR_i)$) can be obtained without having to do a discovery GWAS ourselves[39].

Now, imagine we are conducting a study to examine the effect of an exposure (say, stressful life events (SLE)) on the risk of myocardial infarction (MI). It would be customary to control for known risk factors in our regression so that our model would be written as

$$Y = \alpha + \beta_1 \text{SLE} + \beta_2 \text{smoking} + \beta_3 \text{cholesterol} + \beta_4 \text{diabetes} + \beta_5 \text{hypertension}$$

where Y = case vs. control and the β's are coefficients for each predictor or covariate. We can now add the polygenic score for MI (derived from a prior GWAS) as a covariate:

$$Y = \alpha + \beta_1 \text{SLE} + \beta_2 \text{smoking} + \beta_3 \text{cholesterol} + \beta_4 \text{diabetes} + \beta_5 \text{hypertension} + \beta_6 \text{polygenic score}$$

In this regression, we can estimate the independent effect of the polygenic score on MI, and we can also estimate the independent effect of SLE on MI controlling for risk factors including genetic loading (i.e., polygenic score) for MI.

In general, the power and precision of a polygenic score will increase with increasing size of the discovery sample and by including a larger number of independent variants in the score. A large variety of methods have been developed to calculate polygenic scores that differ mainly in their approach to selecting which DNA variants to include and to assigning weights to the variants.

In recent years, methods that utilize a Bayesian approach have been shown to outperform the classic "clumping and thresholding" approach described above. These methods assume a prior distribution of SNP effect sizes and can incorporate all available SNPs. Special methods are also now available to calculate polygenic scores that incorporate GWAS data from individuals of multiple ancestries[40, 41].

8.16 SNP Heritability and Genetic Correlation

In addition to constructing polygenic scores, GWAS results can be used to calculate the overall heritability of a trait that is captured by common (SNP) variation. This "SNP heritability" (h^2_{SNP}) is analogous to the heritability we calculated from twin studies (Section 8.3) except that we are using directly genotyped SNPs rather than the genetic relationship of twin pairs. Once again, there are many methods available to calculate SNP heritability (see Barry et al.[42] for a more detailed overview of methods), most of which rely on samples of unrelated individuals. Here, we will describe one widely used method known as LD score regression (LDSC)[43]. One advantage of LDSC is that it relies on summary statistics from previously conducted GWAS rather than requiring access to individual-level data.

LDSC makes use of the fact that the effect size of the association between a given SNP and a trait of interest incorporates the effects of all the SNPs in linkage disequilibrium (LD) with (i.e., tagged by) that SNP. Thus, on average, SNPs with greater LD will have higher χ^2 statistics than SNPs with lower LD because their greater extent of LD is more likely to encompass true causal SNPs. We can then calculate an LD score (l_j) for a given SNP, j as:

$$l_j = \sum\nolimits_k r^2{}_{jk}$$

where r^2 is a measure of LD (specifically, it's the squared correlation between alleles of any two SNPs that we get from a reference panel like the 1000 Genomes Project). We get an SNP's LD score by summing the values of r^2 between SNP j and all other (k) SNPs in a given region (e.g., 1 Mb) around j. Thus, the LD score is a measure of all of the genetic variation tagged by SNP_j.

Now, we take all the SNPs included in a GWAS and calculate the expected χ^2 of each SNP given its LD score as:

$$E\left(\chi^2 | l_j\right) = \frac{Nh^2\, l_j}{M} + N_a + 1$$

where N is the sample size and M is the total number of SNPs (and $\frac{h^2}{M}$ is the average heritability per SNP). The N_a term is an adjustment made to account for biases due to population stratification or hidden relatedness of individuals. This is simply a linear regression of the observed χ^2 statistics of our SNPs on their LD scores and can also be written as:

$$E\left(\chi^2 | l_j\right) = \beta_0 + \beta_1 lj$$

Note that LD score regression includes two parameters: the intercept $\beta_{0,}$ which is the same as the N_a + 1 terms, and β_1, which is the slope that we use to estimate heritability:

$$\widehat{h}^2_{SNP} = \beta_1 \frac{M}{N}$$

There are a few points to note about interpreting SNP heritability. First, the estimated heritability can be confounded by population stratification. In the case of LDSC, we can evaluate this bias using the intercept term (β_0): When β_0 is >1, this is evidence that confounding is present[43]. Second, SNP heritability is typically substantially lower than the heritability estimated from twin studies. One important reason for this gap (often referred to as "missing heritability") is that SNP heritability only includes the effects of common variants, whereas, in theory, twin heritability includes all forms of genetic variation. The estimates from LDSC in particular can be lower than those produced by other methods of calculating SNP heritability. Third,

when we estimate the SNP heritability of a binary trait (e.g., using case–control data), we need to apply a correction that treats the trait as though it has an underlying continuous distribution of liability.

In addition to estimating how much of the variance in a trait is due to genetic variation (heritability), we may also want to understand whether and to what extent the genetic basis of one trait (say, coronary artery disease) overlaps with that of another trait (say, type II diabetes). Methods that estimate SNP heritability have also been extended to allow us to estimate the *genetic correlation* between two traits. Returning to the example of LDSC, the genetic correlation (denoted r_g) between two phenotypes can be found by substituting the χ^2 statistics for a single trait with the product of the SNP z scores from two traits. We can then regress this $z_1 z_2$ product on the LD score for each SNP.

$$E\left[z_1 z_2 | l_j\right] = \frac{\rho_g \sqrt{N_1 N_2}}{M} l_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}.$$

Here, N_1 and N_2 are the sample sizes for the GWAS studies of each trait, respectively; ρ_g is the genetic covariance between the traits; M is again the number of SNPs and l_j is the LD score of the jth SNP. We also now have a term $\frac{\rho N_s}{\sqrt{N_1 N_2}}$ that accounts for any overlap of individuals between the two samples, where ρ is the phenotypic correlation between overlapping individuals and N_s is the number of overlapping individuals. The genetic correlation (r_g) between two traits can be calculated from their covariance and their SNP heritabilities:

$$r_g = \frac{\rho_g}{\sqrt{h^2_{SNP_1} h^2_{SNP_2}}}.$$

8.17 Rare Variants and Genome Sequencing

As we have said, GWAS using DNA microarrays are designed to test common SNPs across the genome. However, they do not cover all of the variation that exists in the genome. In fact, common variation represents less than 1% of all the variation in genomes; the rest of the variations are rare. Now, thanks to advances in human genome sequencing, we can directly genotype all variation at relatively low cost. Thus, it has become possible to analyze rare variation in case–control samples.

DNA sequencing studies typically take one of two forms. The first involves sequencing all of the exons (the protein-coding portion of the genome, known as the "exome"). Although the exome makes up less than 2% of the genome, variations in exons are easier to interpret. That's because when a mutation disrupts an exon, it's easy to see how this can lead directly to a disease. The second approach involves

sequencing the whole genome (i.e., "whole genome sequencing"). While this provides much more information than whole exome sequencing, it can be harder to be sure that a change elsewhere in the genome has a functional impact.

As we pointed out earlier, the rare variants detected by sequencing can have much larger effects than the common variants that show association in a GWAS. Variants with large effects are expected to be rare because natural selection will have prevented harmful alleles from becoming common in a population. Since larger effects are easier to detect than smaller ones, you might think that sequencing studies wouldn't need the large sample sizes that are required for GWAS. And that's true. . .sometimes. For example, rare mutations in Mendelian (single-gene) diseases have been discovered by studying a few families that carry the mutation. But for complex diseases like diabetes or schizophrenia—where many different rare variants may be involved, some of which may not even be inherited (de novo mutations)—the sample size issue is still a challenge. That's because rare variants are rare. The power to find a genetic association depends on both the effect size and the frequency of a variant. So, just as we may need many thousands of cases and controls to find a variant that's common but has a small effect (GWAS), we need similar numbers to find a variant that is rare even if it has a large effect.

For rare variants, the standard association tests we use in GWAS won't work because the variants are rare and there are too many of them (potentially hundreds of millions in a whole genome sequencing study). That means we won't have the power to detect a significant effect of any particular rare variant. So, one common strategy is to test groups of rare variants jointly. For example, we can collapse rare variants in a gene and analyze the data at the level of genes. We can then ask whether the "burden" of rare variants is greater among cases than controls after aggregating the variants in each gene into a "burden score." We can then correct for multiple testing by using a Bonferroni correction for the number of *genes* (roughly 20,000) in the genome—i.e., 0.05/20,000. Other strategies for selecting groups of rare variants include collapsing variants in a region or in sets of genes that make up different biological pathways. The variants can also be selected or weighted by their frequency or their likelihood of being functionally important or deleterious (e.g., whether they would disrupt a protein). One pitfall of burden tests is that they assume that all the rare variants are causal and have the same direction of effect (e.g., they all increase risk). If the variants actually are a mixture of risk-increasing and risk-decreasing mutations, they may cancel each other out. A different approach, known as variance component tests, allow for mixed directions of effect. One widely used approach called SKAT-O finds the optimal combination of burden and variance component tests, providing greater flexibility and power when there are a range of effects (risk, protective, and null variants). There are many tests and statistical packages now available for rare variant testing, including those, e.g., STAAR, that can analyze large-scale whole genome sequence data, capable of analyzing noncoding varians and incorporating other biological and functional information to weight rare variants[44].

8.18 Mendelian Randomization and Causality

Genetic data can also be used as a tool for causal inference in epidemiology and biomedicine. It is often said that the gold standard to determine whether a treatment or exposure has a causal effect on an outcome is the randomized, controlled, clinical trial (RCT) (Chapter 6). In a typical RCT, individuals are randomly assigned to one of two treatments or to a treatment or placebo group and the outcomes in the two groups are then compared. If randomization is successful, the two groups being compared will be similar on all other characteristics except the treatment itself, and so we can conclude that the differences we observe in the outcomes are due to the treatment, not to the different composition of the two groups. They will be similar in all these other characteristics because these are randomly (rather than systematically) distributed in the two groups. However, we cannot always do a clinical trial, in which case we rely on observational studies to draw inferences about causality. Although the results of most large and well-designed observational studies are borne out by randomized clinical trials when these are done, there are quite a few observational studies inferring causal effects of certain exposures that are not subsequently confirmed in clinical trials and sometimes even reversed. A dramatic instance of that was the Women's Health Initiative in which the randomized clinical trials of hormone therapy showed harm, thereby upending decades of observational research that showed benefit, presumably due to confounding by the use of hormones with other health behaviors. And there are many instances where RCTs are not possible. For example, to look at the effect of smoking on lung cancer, we would not want to randomly assign people to smoke.

Fortunately, there is a method called Mendelian randomization (MR) that can help us assess causality in circumstances where an RCT is not feasible. For example, we could use MR to evaluate if smoking is a causal factor in lung cancer by using genetic data to randomly assign individuals to higher or lower levels of smoking. We can do this by using genetic variants, which are inherited randomly from parents, and act as an *instrumental variable*—that is, a variable that is correlated with the exposure of interest but not known to be associated with the outcome of interest. What we really want to know is the *causal role* of the exposure on the outcome, and in MR we use the genetic variants as a proxy for the exposure.

For purposes of illustration, and to see the analogy to a clinical trial, let us suppose we have an SNP where the alleles are either A or G and where the A variant has been previously associated with smoking. Thus, on average, people who carry the A allele are more likely to smoke than those with the genotype GG. Which genotype you inherit is a matter of chance, unrelated to any other phenotypic or environmental potential confounders; it is essentially random. Now if you have a large sample, you can think of the two groups as those who carry the A allele, and those who do not, as illustrated in the diagram below (Figure 8.7).

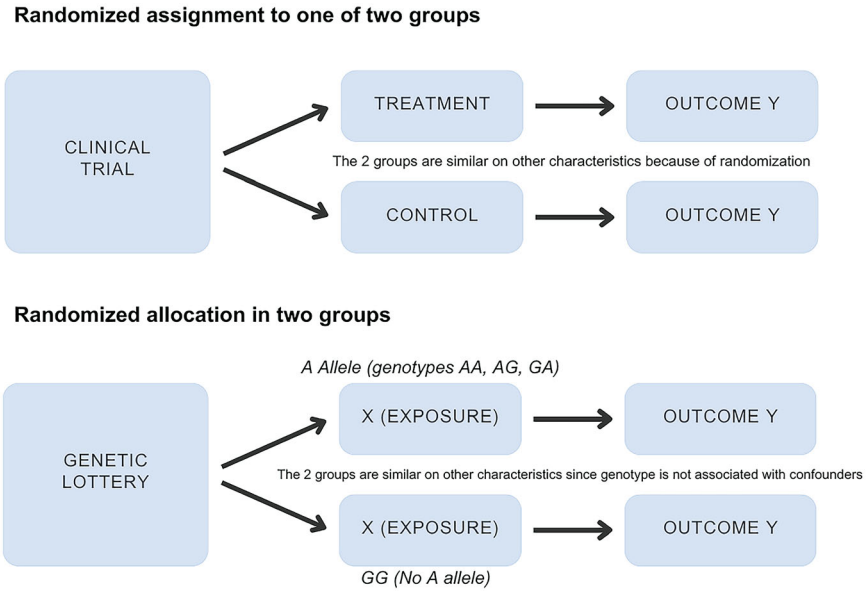


Figure 8.7 Clinical trials versus Mendelian randomization

Thus, people who have the risk allele A have genotypes of AA, AG, or GA, and people who do not have the risk allele have genotype GG. How we can use this to determine causality is described in a section below, but there are certain caveats to selecting the genetic variants to act as instrumental variables. The following assumptions must be met:

(1) Relevance: The genetic variant must be robustly associated with the exposure. (We will be using it as a proxy for the exposure.)
(2) Exchangeability: The genetic variant must not be associated with confounders of the relationship of exposure to outcome.
(3) Exclusion restriction: The genetic variant must be associated with the outcome only through the exposure (i.e., it may not be associated with the outcome through any other pathway than the pathway through the exposure).

We can use the risk allele *count* as the variant score, so that, for example, if the A allele is the risk allele, you would use 0 for no A alleles, 1 for one A allele, and 2 for both AA alleles being present. You can also use a combination of SNPs rather than just one SNP. For instance, you might devise a score composed of all the SNPs that are genome-wide significantly associated with your exposure in prior GWAS and treat the score as a continuous predictor or compare outcomes between a group with the highest scores to one with the lowest scores.

Figure 8.8 shows the three key elements of our Mendelian randomization analysis: V (the A/G genetic variant or "instrumental variable), X (our exposure of interest, say smoking), and Y (the outcome of interest, say lung cancer).

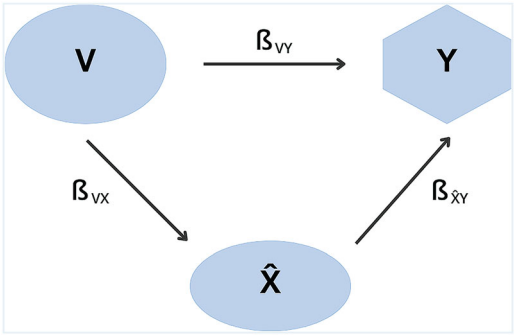


Figure 8.8 Key elements in Mendelian randomization

8.18.1 Two-Stage Least Squares Estimation

Mendelian randomization may be used to estimate causal effects in observational studies through a sequential two-step process. First, we look at the relationship (β_{vx}) between the genetic variant and the exposure X, and in the second step, we look at the relationship (β_{vy}) between the variant V and the outcome Y as shown in the diagram above. What we are ultimately interested in is $\beta_{\widehat{XY}}$—that is the association beween the exposure and the outcome.

If the variant satisfies the MR assumptions, then this estimated exposure $\widehat{X}$ is independent of confounders and so the relationship between $\widehat{X}$ and Y, i.e., $\beta_{\widehat{XY}}$, can be considered to be causal.

A common way to estimate the causal effect $\beta_{\widehat{XY}}$ is to note that the direct effect of the variant on the outcome, given by β_{VY}, is the product of the effect of V on X multiplied by the effect of X on Y (see Section 4.28):

$\beta_{VY} = (\beta_{VX})\,(\beta_{\widehat{XY}})$; therefore, the causal effect of X on Y is estimated as:

$$\beta_{\widehat{XY}} = \frac{\beta_{VY}}{\beta_{VX}}.$$

8.18.2 Wald Test Statistic

The Wald test of the significance of the causal effect of X on Y is $\beta_{\widehat{XY}}/\mathrm{SE}\,(\beta_{\widehat{XY}})$, i.e., it is the ratio $\frac{\beta_{VY}}{\beta_{VX}}$ divided by the standard error of the ratio. If the p value of this test statistic is less than 0.05, we can reject the null hypothesis and conclude there is a significant causal relationship (after accounting for multiple testing, if any).

To test the significance of $\beta_{\widehat{XY}}$, we need to get the standard error of $\beta_{\widehat{XY}}$, which is the square root of the variance of $\beta_{\widehat{XY}}$. An estimate of the variance of $\beta_{\widehat{XY}}$ is:

$$\text{Var}\ \beta_{\widehat{XY}} = \text{Var}\ \frac{\beta_{VY}}{\beta_{VX}} = \text{Var}\ \frac{(\beta_{VY})}{\beta^2{}_{VX}} + \frac{\beta^2{}_{VY}}{\beta^4{}_{VX}}\ \text{var}\ (\beta_{VX}) - 2\frac{\beta_{VY}}{\beta^3{}_{VX}}\ \text{covariance}\ (\beta_{VY},\ \beta_{VX})$$

If β_{VY} and β_{VX} are estimated from different samples, the covariance is 0 and the last term disappears.

8.18.3 Caveats

As we noted above, we can only use MR if we can satisfy the three assummptions described earlier. The relevance assumption means that we the genetic variant (s) must be robustly associated with and account for more than a small proportion of the variance of the exposure (X). Variant(s) that do not fulfil this criterion are called "weak instruments." One of the consequences of weak instruments is that the standard errors for the betas are large and there may not be enough power to detect a true causal effect.

A second challenge with MR is that it is often difficult to find an appropriate instrumental variable, i.e., a genetic variant, or a combination of variants, that we can be confident is related to the outcome only through the exposure (the exclusion restriction criterion). And that is because we know that there is a lot of *pleiotropy* in human genetics. Pleiotropy means that a variant or gene affects multiple traits or phenotypes. So if the instrumental variable we choose is a genetic variant that also affects the outcome Y directly (not just through the exposure X) that is related to the outcome, we have violated the exclusion restriction assumption. To address this, we can use statistical methods that ameliorate this problem, like MR-EGGER and MR-PRESSO. Explanations of these methods can be found in the "References" section for this chapter[45].

Neverthess Mendelian randomization is being increasingly used and new statistical methodologies continue to be developed to evaluate and address violations of the MR assumptions and for improving estimates of causality. The summary of MR provided here is the tip of the iceberg, touching on the most commonly used methods, but different approaches apply to different research problems and datasets.

There has been a paradigm shift in science—from believing things are simpler than they seem to understanding they are more complex than they seem. For the last century, the principle guiding scientific endeavor was Occam's razor—that the most parsimonious explanation for phenomena is the best. But as genomic and molecular discoveries accelerate, it becomes apparent that in the biological sphere, simple explanations are not possible and the aim is to more accurately uncover and explain the inherent complexity (and marvel) of life.

Chapter 9
Risk Prediction and Risk Classification

It is far better to foresee even without certainty than not to foresee at all.
Henri Poincare in The Foundations of Science, p. 129.

9.1 Risk Prediction

We are interested in predicting risk of a disease for an individual because treatment decisions are often based on risk. We will use cardiovascular risk as an example in the following sections since risk prediction is most developed for this disease. Treatment guidelines for high blood pressure or high cholesterol from the American Heart Association differ for people at high risk of cardiovascular disease from those at lower risk. For example, anticoagulant drugs are recommended to people who have atrial fibrillation (a type of heart arrhythmia) if they are at high risk of stroke. Since anticoagulants pose a risk of bleeding, they are not recommended for people who have a low risk of stroke. Prediction of risk is also very useful for public health matters. Knowing what percentage of a population is at high risk can help health planners to mount preventive measures and to plan utilization of resources.

Traditionally accepted risk factors for stroke generally available to clinicians are age, systolic blood pressure, diabetes mellitus, cigarette smoking, prior cardiovascular disease, atrial fibrillation, left ventricular hypertrophy by electrocardiogram, and the use of antihypertensive medication. Equations estimating risk for stroke were developed from prospective studies like the Framingham study (see Section 4.11) in which a specified population in the town of Framingham Massachusetts had measures of these variables and were followed up over time to see who developed events like stroke or heart attack and how well the baseline variables studied could predict who would develop the event or outcome[46–48]. With the advent of more sophisticated techniques of measuring certain proteins in the blood (biomarkers), people became interested in refining risk prediction models and in seeing whether certain biomarkers could improve prediction. The question is how do we evaluate whether or how much the new biomarker improves risk prediction? It is an important question because if the biomarker does help in prediction, it may become routinely used in doctors' offices. Adding biomarkers to risk prediction may indicate that some

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_9

people, previously thought to be at low risk, are at higher risk and should be treated. On the other hand, this may add to health-care costs, so if it does not improve prediction and has no bearing on treatment decisions, it may not be worth doing the test routinely.

The first indication that a biomarker may be useful is that it is significantly associated with the outcome of interest, either in a logistic regression or a Cox proportional hazards regression model (see Section 4.20 for explanation of Cox proportional hazards models). The important point to note is that a biomarker may be significantly associated with a disease outcome but that doesn't mean it necessarily improves risk prediction or reclassifies people into different categories of risk.

As an example, let us consider a protein called C-reactive protein (CRP), which is a marker of general inflammation and can be assayed from a simple blood test. We want to see if it adds to the prediction of stroke.

The ideal situation would be if we measured the biomarker of interest in everyone before the event and then followed them forward in time as some developed the outcome and others did not (in other words, a prospective study) and then added the biomarker to the variables in our risk prediction model. But it gets very expensive to do that, so the usual approach is to do a case–control study and measure the biomarker only in the cases and their controls. This chapter discusses the measures we use to evaluate the additive value of a biomarker in predicting risk.

9.2 Additive Value of a Biomarker: Calculation of Predicted Risk

We will consider risk prediction of ischemic stroke as our example, using data from a case–control study of ischemic stroke nested within the prospective Women's Health Initiative Observational Study of postmenopausal women.[49] We consider the biomarker C-reactive protein (CRP), a marker of inflammation. There were 868 cases and 883 controls who had CRP assayed at the baseline examination and who had had no history of prior stroke. (Actually, the 883 stroke cases were matched to 883 controls on several variables, but 15 of the cases did not have an adequate blood sample, leaving 868 cases and the 883 controls.)

First we need to develop a risk prediction equation from our data and then we need to estimate the probability of stroke during a specified period of time for each person in our study population using the risk prediction equation without the biomarker in it. This probability will depend on the specific values for each person of the variables we use in the prediction model. The general idea is that first we have to classify people into risk categories based on our old model without CRP. Next we classify the same people based on our new model, which consists of the variables in the old model plus the biomarker CRP. We compare the two classification schemes with the actual outcomes and see which model predicts more accurately. The

sections below describe several measures used to evaluate how well the biomarker adds to prediction. Some excellent papers by Pencina go into further detail[50–52].

We first run a logistic regression model from which we can calculate the probability of stroke. We use variables traditionally used in the prediction of stroke in the Framingham study: age, systolic blood pressure, diabetes mellitus, cigarette smoking, prior cardiovascular disease, atrial fibrillation, left ventricular hypertrophy by electrocardiogram, and the use of antihypertensive medication in an equation to predict the future occurrence of stroke. (In our example we left out LVH because we did not have data on that and because the prevalence of LVH in the WHI population was very low. Furthermore, the use of antihypertensive medication is a variable that depends on the level of blood pressure, but for simplification we just consider medication = 1 if the woman uses it and = 0 if she does not.)

Recall from Chapter 4 that in a logistic regression, the probability of an event is

$$P(event) = \frac{1}{1 + e^{-k}}$$
$$where\ k = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \ldots + C_mX_m$$

The parameters in k (the values of C_0 to C_m) are obtained from the logistic regression that can be run in various computer statistical packages like SAS or STATA. In our stroke case–control study, the unconditional logistic regression we calculated with the variables used in the Framingham risk score, *k*, was

$k = (-2.4421 - 0.0108$ (age) $+ 0.0207$(SBP) $+ 0.0003$ (on blood pressure medication) $+ 0.2228$ (history of CHD) $+ 0.9868$ (current smoker) $+ 0.5829$ (atrial fibrillation) $+ 0.7867$ (diabetes) $+ 0.2411$ (Caucasian))

But we must make an adjustment to *k* to account for the fact that it is a case–control study. Thus, we add the term $ln\left(\frac{P}{1-P} \cdot \frac{n_{controls}}{n_{cases}}\right)$ to the intercept of the unconditional logistic regression model; P is the probability of stroke in the target population. In a prospective study, we can get P directly because it will just be the number of stroke cases over our follow-up time period (we are choosing 8 years). But because this is a case–control study and P by definition is 50%, i.e., there are 50% stroke cases and 50% controls in our design, we have to estimate P for the target population. For WHI we estimate P as the annual incidence of stroke in the WHI Observational Study, which was 0.0029 annually, times the average follow-up of 8 years or 0.0232. Thus, our correction factor is

$$\text{correction factor } \ln\left(\frac{P}{1-P} \cdot \frac{n_{controls}}{n_{cases}}\right) = ln\left(\left(\frac{.0232}{.9768}\right) \cdot \left(\frac{883}{868}\right)\right)$$
$$= \ln(.02375 \ times\ 1.01728) = \ln\ (.02416) = -3.7230$$

Note if there is an equal number of cases and controls, then $\frac{n_{controls}}{n_{cases}}$ becomes 1 and the correction factor is simply the natural log of $\frac{P}{1-P}$

Our corrected $k_{corrected} = k +$ correction factor

Next we calculate the probability of stroke within 8 years for each person in our study, using that person's values of the variables. So, for example, for the i_{th} person who is a white, nonsmoking woman of age 55, with systolic blood pressure of 120, who does not have diabetes, is not on antihypertensive medication, with no atrial fibrillation or history of heart disease, the probability of stroke in 8 years is calculated as follows:

$k_{\text{corrected}} = (-2.4421 - 0.0108\,(55_{\text{age}}) + 0.0207(120_{\text{sbp}}) + 0.0003\,(0_{\text{medications}}) + 0.2228\,(0_{\text{chd}}) + 0.9868\ \ (0_{\text{current smoker}}) + 0.5829\ \ (0_{\text{atrial fibrillation}}) + 0.7867\ \ (0_{\text{diabetes}}) + 0.2411\,(1_{\text{Caucasian}}) + (-3.7230_{\text{correction factor}}) = -4.0340$

and $P_i = \frac{1}{1+e^{-k}} = \frac{1}{1+56.4862} = \frac{1}{57.4862} = .0171$

So that person has an estimated risk of stroke in 8 years of 1.7%.

We calculate these probabilities for each of the stroke cases and each of the controls. Next we divide these probabilities into risk categories. We have chosen the following risk categories: <2%, 2% to <5%, 5% to <8%, and ≥8%. The usually accepted categories are <5% low risk, 5 to less than 10% low intermediate risk, 10–20% high intermediate risk, and >20% high risk. But in our example, we have a generally low to intermediate risk population (by virtue of the fact that we excluded all those with a previous stroke or heart attack). The risk categories we chose roughly correspond to low, intermediate, and high risk levels used in decisions to initiate treatment to prevent stroke in persons with atrial fibrillation. The person in our example above is in the lowest risk category of <2%.

We repeat this same process using the new model that includes the variable CRP. Note that this new model with the CRP has slightly different coefficients for each of the variables than did the old model without the CRP. (We use lnCRP—the natural log of CRP—because CRP is not normally distributed.)

Let us assume our sample woman has a CRP level of 12 (which is very high):

$$\ln(12.0) = 2.4849$$

In our example, $k_{CRP,\ corrected} = (-2.7246 - 0.0078(55_{\text{age}}) + 0.0201(120_{\text{sbp}}) + 0.0000(0_{\text{medications}}) + 0.2048(0_{\text{history of chd}}) + 0.9789(0_{\text{current smoker}}) + 0.5615(0_{\text{atrial fibrillation}}) + 0.7003(0_{\text{diabetes}}) + 0.2076(1_{\text{Caucasian}})) + 0.1954\ \ (2.4849_{\text{lnCRP}}) + (-3.7230) = -3.7714$ and her estimated probability of stroke in 8 years is shown below:

$$P_{i(\text{CRP})} = \frac{1}{1+e^{-k}} = \frac{1}{1+44.4746} = \frac{1}{45.4746} = .0215$$

or a 2.2% probability of a stroke in 8 years.

Adding her CRP level to the model increased her low risk from 1.7% to 2.2%. Thus, she has moved from the lowest risk category (<2%) to a higher risk category (2% to <5%). This may or may not lead her doctor to change her treatment, but it may lead the doctor to do more active surveillance of her risk factors. Doing these same calculations for a 65-year-old, Caucasian, diabetic women with a systolic blood pressure of 170 whose risk with the old model was 8.2% and whose CRP was 5.0 (still higher than the 3.0 considered in the normal range, but not as high as

the CRP of 12 in our previous example), does not really change her risk much, which is now 8.3%. Since she was high risk to begin with, though her CRP was higher than normal but not terribly much so, she remains at high risk and the other risk factors, like her diabetes and high blood pressure, overwhelm the effects of the elevated CRP. Now of course, what we want to know is are these women cases or controls? We have to test our predictions against the actual outcomes.

After calculating the risk with and without CRP for all women in the study, we next cross-tabulate the risk categories for the two models separately for those who had a stroke during the 8-year period and those who did not have a stroke to get the table below:

Predicted risk with Old model with no CRP

Predicted risk with the new model thatincludes CRP

Stroke Cases

Risk categories	<2%	2 to <5%	5 to <8%	>=8%	total
<2%	**6**	3	0	0	9
2 to <5%	20	**336**	67	0	423
5 to <8%	0	44	**157**	42	243
>=8%	0	0	20	**173**	193
total	26	383	244	215	868

Controls

Risk categories	<2%	2 to <5%	5 to <8%	>=8%	total
<2%	**23**	10	0	0	33
2 to <5%	63	**480**	51	0	594
5 to <8%	0	44	**107**	26	177
>=8%	0	0	11	**68**	79
total	86	534	169	94	883

Note that the numbers on the diagonal represent women whose predicted risk was not changed by the addition of CRP to the model. For example, there were 336 stroke cases with predicted risk of 2% to <5% with both the old and the new model and 480 of the no-stroke controls that remained in the same risk category. The numbers above the diagonal represent people whose risk went up with the addition of the biomarker to the model and those below the diagonal represent people whose predicted risk went down with the new model that includes the biomarker.

What we would like to see with a new biomarker is that stroke cases should go up in a predicted risk when the biomarker is added and no-stroke controls should go down in a predicted risk. In other words, we want the predicted risk to move in the *right direction*: Stroke cases should move up and controls should move down. From our example we have the following: (proportion (up|case) means proportion moving

up in risk among cases; proportion (down|control) *means* proportion moving down in risk among controls). Thus, 26.3% of our group were reclassified in the right direction with the addition of CRP and 19.6% were reclassified in the wrong direction, for an improvement in classification of 26.3%–19.6% or 6.7%.

Prop(up|case)	$\frac{\#\text{cases moving up}}{\#\text{cases}}$	$\frac{112}{868} = .129 = 12.9\%$	Right direction
Prop (down|case)	$\frac{\#\text{cases moving down}}{\#\text{cases}}$	$\frac{84}{868} = .097 = 9.7\%$	Wrong direction
Prop (up|control)	$\frac{\#\text{controls moving up}}{\#\text{controls}}$	$\frac{87}{883} = .099 = 9.9\%$	Wrong direction
Prop (down|control)	$\frac{\#\text{controls moving down}}{\#\text{controls}}$	$\frac{118}{883} = .134 = 13.4\%$	Right direction

9.3 The Net Reclassification Improvement Index

NRI is = (prop (up|case) – prop (down|case)) – (prop (up|control) – prop (down|control))

In our example it is $(0.129 - 0.097) - (099 - 0.134) = 0.032 - (-0.035) = 0.032 + 0.035 = 0.067 = 6.7\%$

This means that by adding the biomarker, we have correctly reclassified 6.7% of people (i.e., had a net improvement in classification). Another way to write this is below, where $\widehat{p}$ is the estimated probability (which we estimate from our sample by looking at the proportion), D = D is disease (or stroke) = 1 and D = 0 no disease (or no stroke) = 0:

$$NRI = [\widehat{p}(up|D=1) - \widehat{p}(down|D=1)] + [\widehat{p}(down|D=0) - \widehat{p}(up|D=0)]$$

But is this significantly different from zero? To find out, we need to calculate a z score (see Chapter 3) which is the NRI divided by its standard error. The formula is given below.

$$z = \frac{NRI}{\sqrt{\frac{\widehat{p}(up|D=1) + \widehat{p}(down|D=1)}{\#events} + \frac{\widehat{p}(up|D=0) + \widehat{p}(down|D=0)}{\#non-events}}}$$

In our example,

$$z = \frac{0.67}{\sqrt{\frac{.129+.097}{868}+\frac{.099+.134}{883}}} = \frac{.067}{\sqrt{.000260+.000264}} = \frac{.067}{\sqrt{.000524}} = \frac{.067}{.02289}$$

$$= 2.93$$

Z of 2.93 has a p value of <0.01; thus, the NRI is significantly different from 0.

The NRI is one index of the additive value of a biomarker to risk prediction. A drawback, however, is that it depends on the absolute risk categories we select. The category-less NRI is another index, which does not depend on how we categorize risk.

9.4 The Category-Less NRI

The category-less NRI just looks at movement in predicted risk up or down when the biomarker is in the model, regardless of whether this movement means people crossed over to another risk category. It is the percent of all subjects whose risk estimates are changed in the correct direction minus the percent changed in the incorrect direction to get a net effects figure. The correct direction for cases is increased risk in the model with the biomarker compared to the model without the biomarker, and for controls it is decreased risk in the model with the biomarker compared to the model without the biomarker. So for each person, we have to calculate the probability of stroke using her values of the variables in the model with and without the biomarker, and then we count.

It is calculated as

$$\frac{\textit{number of cases whose risk with biomarker is greater than risk without the biomarker}}{\textit{total number of cases}}$$
$$+\frac{\textit{number of controls whose risk with biomarker is less than risk without the biomarker}}{\textit{total number of controls}}$$
$$-\frac{\textit{number of cases whose risk with biomarker is less than risk without the biomarker}}{\textit{total number of cases}}$$
$$-\frac{\textit{number of controls whose risk with biomarker is greater than risk without the biomarker}}{\textit{total number of controls}}$$

Basically, we see what proportion of people had their risk changed in the correct direction (up for cases, down for controls) and subtract the proportion of people who had their risk changed in the wrong direction (down for cases and up for controls).

In our example, the category-less net reclassification improvement was 18.9%, meaning that 18.9% of people had their risk estimate changed by any amount in the right direction when CRP was added to the prediction model.

9.5 Integrated Discrimination Improvement (IDI)

A quality we want in adding a new biomarker to a prediction equation is that it should improve discrimination between those who suffer a stroke from those who don't. An index of discrimination is the integrated discrimination improvement (IDI) measure, where we compare average predicted probabilities for cases and controls. To calculate IDI we first calculate two quantities for each person: (1) the probability of stroke for each person using the logistic regression model *without* CRP in it (old model) and (2) the probability of stroke using the model *with* CRP in it (new model). We then get the average of these probabilities for the stroke cases and the average for the controls. IDI is the difference in mean predicted probabilities between cases and controls using the new model minus the difference in mean predicted probabilities between cases and controls using the old model.

We denote $\widehat{p}$ = estimated probabilty

$$\begin{aligned}\text{IDI} = &\left[\text{mean}\,\widehat{p}_{(\text{new model,cases})} - \text{mean}\ \widehat{p}_{(\text{new model,controls})}\right] \\ &- \left[\text{mean}\ \widehat{p}_{(\text{old model,cases})} - \text{mean}\ \widehat{p}_{(\text{old model,controls})}\right]\end{aligned}$$

IDI for our example $= (0.0688_{\text{new model, cases}} - 0.0474_{\text{new model, controls}}) - (0.0661_{\text{old model, cases}} - 0.0475_{\text{old model, controls}}) = 0.0028$ We can test the null hypothesis that IDI = 0 by calculating Z (see Chapter 3) as

$$Z = \frac{IDI}{\sqrt{SE^2(\text{cases}) + SE^2(\text{controls})}}$$

To get SE (cases), we get the differences for each person between the predicted probability with the old model and the new model. Then we get the standard error of these differences. (Recall that $SE = \frac{standard\ \ deviation}{(\sqrt{n}\ \ )}$.) The SE (controls) is calculated in the same way but using the paired differences in probabilities in controls.

In our example the IDI of 0.0028 is statistically significant with $p < 0.001$.

9.6 C-Statistic

Another measure of discrimination is the c-statistic, which is the probability that a randomly selected person with the event of interest (stroke in our example) will have a higher predicted risk than a randomly selected person without the event. The c-statistic reflects the area under the curve (AUC) from a receiver operating characteristic (ROC) curve (see Section 5.2). (The ROC curve is a plot of sensitivity on Y axis vs. 1-specificity on x-axis). Higher c-statistic values indicate better discrimination. Generally, the c-statistic does not move very much with the addition of a

biomarker and is somewhat hard to interpret in this context. Nevertheless, it is widely used and reported even though it may not be a very good measure of the added predictive value of a new biomarker. SAS will print out the c-statistic when running a logistic regression analysis as will STATA.

9.7 Caveats

Note that the prediction equation you get from your own data will always predict better than when applying it to some other population. So you should validate it on another sample. Or if your dataset is large enough, you can do internal validation by developing the model on half the sample and testing it on the other half.

Note also that a new biomarker is clinically useful if it results in some change of treatment or some other action, like increased surveillance. If no treatment decisions will be affected by improving your risk prediction, then you might as well use the old variables on which to base your treatment decisions, though the new biomarker might still be interesting scientifically.

9.8 Summary

To evaluate the added predictive ability of a new biomarker, we take the following steps:

1. Ascertain whether the biomarker is significantly associated with the outcome of interest (from the literature or from your own study).
2. Develop a risk prediction model using traditionally accepted variables but without the biomarker. Use Cox proportional hazards regression models if you have prospective data. Use logistic regression models with a correction factor if you have a case–control study.
3. See if the c-statistic for the new model with biomarker increases significantly from the old model without the biomarker (Section 9.6).
4. Calculate probabilities of stroke for each person from the model with and without the biomarker.
5. Calculate the NRI (see Section 9.3).
6. Calculate the category-less NRI (Section 9.4).
7. Calculate the IDI (Section 9.5).
8. Report all of these measures as no one measure is perfect.

Chapter 10
Research Ethics and Statistics

Morality, like art, means drawing a line someplace.
Oscar Wilde (1854–1900)

10.1 What Does Statistics Have to Do with It?

At first glance it may seem that statistics and research ethics have nothing to do with each other. Not so! Consider why so many people volunteer for medical research studies. In many cases it is because there is an expected benefit. For example, in cancer clinical trials, often the investigational drug is a last hope and may not be available outside of the trial. In many cardiovascular disease studies, participants appreciate the additional care and attention and are willing to try a new drug, for example, for hypertension. And in fact, it has been shown that often, clinical trial participants live longer and do better than the general population even if they are treated with a placebo. But what is perhaps not sufficiently appreciated is that many, many people participate in studies out of altruism to advance scientific knowledge. Scientific knowledge is not advanced when a study is poorly designed, or carried out without sufficient rigor, or not large enough to give an answer. Proper statistics are a determinant of the ethics of a study.

A prime example is the Women's Health Initiative (WHI), described in Chapter 6. Postmenopausal women were asked to join a study of hormone replacement therapy; the study would continue for up to 12 years before the results were known and might not directly benefit the women themselves, but they would answer the important question of the effect of hormones on cancer, heart disease, and osteoporosis. Many of the WHI participants took part for their daughters and granddaughters, and they expressed pride and enthusiasm for answering the questions for future generations. And indeed they did achieve that goal—one part of WHI, the estrogen plus progestin trial versus placebo, has already answered these important questions, with a startling result: Estrogen plus progestin increases risk of breast cancer and also increases heart attacks, stroke and blood clots, and dementia. So although the treatment does show benefit with regard to colorectal cancer and osteoporotic fractures, the overall risks outweigh the benefits. This trial has changed medical practice for generations to come.

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2_10

Well that brings us back to statistics. This study was able to answer these questions because it had sufficient power to answer them. It required 16,608 women in that part of WHI to be able to detect these effects. Even if the result had been null (i.e., if it showed no difference between the treatment and placebo groups), we could have had faith in that result because the power was there to detect a true effect if there really was one. As it turned out, the results were clear-cut, though unexpected, in favor of placebo. So the point is that in order for a study to be "ethical," it must be designed and powered well enough so that it can answer the questions it poses. Otherwise, people who consent to participate in the expectation that they will contribute to knowledge may actually not be contributing because the study is poorly designed, powered, or executed and may be needlessly exposed to risk.

Note that there are certain study designs for which power considerations are less relevant. Examples are pilot studies, which by definition are intended to test the feasibility of a research protocol or to gather preliminary data to plan a full study and are exempt from the power issue. Power considerations may also not apply to certain drug toxicity studies (Phase I trials) or certain types of cancer trials, but certainly in prevention trials, as well as Phase III treatment trials, power is a major consideration in the ethics of research.

10.2 Protection of Human Research Subjects

Human subjects in medical research contribute greatly to improving the health of people. These volunteers must be protected from harm as much as possible. In the not-too-distant past, there were some egregious breaches of ethical principles in carrying out medical research. The world's most appalling examples are the medical experiments carried out in the Nazi concentration camps—by doctors! It defies any kind of understanding how educated, presumably "civilized" professionals could so have distorted their profession and their own humanity. But these atrocities did occur and demonstrate the horrors people are capable of perpetrating. When these atrocities became known after World War II, the Nuremberg trials of Nazi war criminals (including the doctors who preformed such research) also resulted in the Nuremberg Code for conduct of medical research that established the basic requirement of voluntary informed consent. Subsequently, the Declaration of Helsinki, in 1964, expanded and refined the research guidelines and became a world standard, which undergoes periodic revisions.

The most infamous example of unethical research in the United States was probably the Tuskegee Institute study of syphilis, which took place in the south in the United States from 1932 to 1972. The researchers wanted to study the natural course of syphilis. In the 1940s antibiotics became available that could treat this disease, but were withheld from the participants, who were poor Black men, because an intervention to treat the disease would interfere with this observational study. In 1972 the public became aware of this experiment and in 1974 the National

Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established. They developed a report known as *The Belmont Report:* Ethical Principles and Guidelines for the Protection of Human Subjects of Research. These guidelines are followed by all medical schools and other research institutions that conduct research involving human participants, and they are deemed to be universal principles, cutting across cultural lines. The guidelines are based on three basic principles: *respect for persons*, *beneficence*, and *justice*.

Respect for persons recognizes that people are autonomous beings and can make their own informed choices about participating in research, free of coercion. The informed consent process is predicated on this principle. Participants who are not able to make their own choices, such as comatose patients, or mentally incapacitated persons, or young children, must have special protections.

Beneficence, or the principle of non-malfeasance, means that the risks of the research must be kept to a minimum, the benefits maximized, and the researcher is responsible for protecting the participant.

Justice in this context refers to a fair distribution of the risks and benefits of research. One group of people should not be exposed to research risks for the benefit of another group of people. This can get to be a pretty complicated concept. While it may be easy to discern breaches in certain situations—to take the most extreme example, prisoners of the Nazis were subjected to freezing experiments to benefit soldiers who might have to fight under arctic conditions—it may be more subtle in many situations and these must be examined carefully, according to this principle.

10.3 Informed Consent

One of the most important elements in protection of human subjects is the principle of informed consent. The study subject must freely consent to be part of the study after being fully informed of the potential risks and benefits.

There are certain elements that must be in a written consent form. The purpose of the research must be stated; a 24-h contact person must be listed; there must be a description of the study procedures: and what is expected of the participant, the duration of the study, and how much of the participant's time it will take. The potential risks and discomforts, potential benefits, and inconvenience to the participants must all be clearly stated. There must be a statement that participation is voluntary and that the participant has the right to withdraw at any time and that this will not prejudice the care of the participant. If the research may result in need for further care or diagnostic procedures, the participant must be told to what extent he or she is responsible for further care and what the study will pay for. If there is any compensation to the participants, either for expenses incurred in participating or time spent, they must be informed of the amount. (The amount should not be excessive, as that may appear coercive.) A statement assuring confidentiality and how it will be maintained must be included.

Most importantly, the participant must understand what he or she is agreeing to and the consent form must be phrased in language that is understandable and, if appropriate, translated into the participant's native language. All this must be approved by the medical institution's Institutional Review Board (IRB), which is generally a committee of experts and lay people who review and must approve all research protocols before the research is started and who monitor adverse events as the research progresses. Different IRBs have different specific requirements that are usually posted on their websites. Informed consent is an ongoing process—it is not just signing a form at the beginning of a study. The researcher has an obligation to keep the participant informed of relevant new research that may affect his or her decision to continue participating.

Back to the WHI, since it was believed at the time WHI was started that hormones would protect women from heart disease, the initial consent form stated this as a potential benefit. Potential risks stated in the consent form included an increase in breast cancer and blood clots. When WHI was in progress, the Heart and Estrogen Replacement Study (HERS) published results indicating that for women who already had heart disease (secondary prevention trial), hormone replacement provided no benefit. They observed more heart attacks in the early part of the study, with a possible late reduction, resulting in no overall difference between the treatment and placebo groups by the end of the study. This information was provided by a special mailing to all women participating in the WHI hormone program for primary prevention of heart disease. (Primary prevention means the study was carried out in generally healthy women.) Subsequently, early data from the WHI itself indicated there was early harm with respect to heart disease. Again, the women were informed by a special mailing, telephone, and personal discussion with clinic staff. Ultimately, the estrogen plus progestin trial was stopped after 5.2 years (instead of the originally planned average of 8.5 years) because the excess breast cancer risk crossed a predetermined stopping boundary and a global index of overall effects suggested more harm than benefit, and all women in the trial were discontinued from their study pills.

10.4 Equipoise

That brings us to another concept: When is it ethical to begin a clinical trial of a new treatment? When there is equipoise. *Equipoise* means that there is about equal evidence that the treatment may provide benefit as there is that it will not provide benefit. If we are sure the treatment provides benefit, we should not deny it to people who would be getting placebo in the trial. Of course we may be wrong. There were critics of the Women's Health Initiative who said it was unethical to do such a trial because it was well known that hormones protect against heart disease and it would be unethical to deny these hormones to the women randomized to placebo! Of course we now know that was wrong—the placebo group did better. At the time WHI was started, the observational evidence pointed to benefit with regard to heart disease, but

it had never been tested in a clinical trial, which is the "gold standard." Thus, there were many people who did not believe that the benefits of hormone replacement were already established by the observational studies, and it turns out they were right. The researcher, whose obligation is to protect human research participants, must believe it is equally likely that the treatment is better or that the placebo or comparison treatment is better. The scientific community that judges the research proposal must believe, based on the "state of the art," that there is a reasonable question to be answered.

10.5 Research Integrity

For research conclusions to be valid, data collection procedures must be rigorously and uniformly administered. No data may be altered without documentation. If there is a clerical error, the change and reason for it must be documented. Enrollment must be according to strict and preplanned standards. Sometimes (fortunately, rarely) there is a great pressure to enroll subjects in a given time frame, or the researcher (in violation of the principle of equipoise) really believes the treatment can help his or her patients and so "bends" the enrollment rules. This may invalidate the research and so is unethical. A very sad example of this occurred in the National Surgical Adjuvant Breast and Bowel Project (NSABP). This multicenter study demonstrated that lumpectomy could be equivalent to mastectomy in hundreds of thousands of women. The chairman of this study discovered that the principal investigator in one of the clinical centers had falsified some patient records so that women who were not eligible to be in the study based on predetermined enrollment criteria were made falsely eligible to participate. This excellent and extremely important study was initially tainted when this became known and the chairman of the study was charged by the Office of Research Integrity (ORI) with scientific misconduct, even though he had notified the NIH of the problem when he learned of it. He was subsequently completely cleared, and he was offered multiple apologies. The study has had profound implications on the treatment of women with breast cancer. Nevertheless, this was a serious breach of ethics on the part of an investigator in one of the many centers that could have invalidated the findings. Fortunately the results held up even when all the patients from the offending clinic were excluded.

10.6 Authorship Policies

In medical research most original research articles have multiple authors, since medical research is a collaborative effort. Most medical journals, and research institutions, have specific and strict authorship policies (published in journals and/or on websites), many of which embody the following elements: (1) Coauthors must make an intellectual contribution to the paper (e.g., conceive the research,

perform analyses, write sections of the paper, or make editorial contributions), (2) all coauthors must bear responsibility for its contents, and (3) coauthors must disclose potential conflicts of interest (e.g., relevant support from industry, lectureships, stock ownership). Order of authorship may sometimes be a point of contention and should be discussed by the coauthors early in the process.

10.7 Data and Safety Monitoring Boards

Generally, clinical trials have a Data and Safety Monitoring Board (DSMB) to oversee the trial. These are independent groups of experts in the relevant disciplines who are in an advisory capacity. Their job is to monitor the trial and to assure the safety of participants. In a blinded trial, they are the only ones who see the unblinded data at regular, prespecified intervals. If they find excessive benefit or harm in one arm of the trial, they would advise to stop the trial (as happened in the Women's Health Initiative). Usually the criteria for stopping a trial due to harm in the treatment group are more stringent than stopping for benefit.

10.8 Summary

The ethical conduct of research has many components. New and difficult ethical questions arise as science advances and new technologies become available. This brief chapter just begins to give you an idea of some of the issues involved. Much more detailed information is available from various websites and NIH has an online course in protection of human subjects. Local IRBs can give you information and additional sources.

Postscript

A Few Parting Comments on the Impact of Epidemiology on Human Lives

Years ago a woman with breast cancer would be likely to have a radical mastectomy, which in addition to removal of the breast and the resulting disfigurement would also include removal of much of the muscle wall in her chest and leave her incapacitated in many ways. Today, hardly anyone gets a radical mastectomy and many do not even get a modified mastectomy, but, depending on the cancer, may get a lumpectomy that just removes the lump, leaving the breast intact. Years ago, no one paid much attention to radon, an inert gas released from the soil and dissipated through foundation cracks into homes. Now it is recognized as a leading cause of lung cancer. The role of nutrition in prevention of disease was not recognized by the scientific community. In fact, people who believed in the importance of nutrients in the cause and cure of disease were thought to be faddists, just a bit nutty. Now it is frequently the subject of articles, books, and news items, and substantial sums of research monies are invested in nutritional studies. Such studies influence legislation, for example, the regulations that processed foods must have standard labeling, easily understood by the public at large, of the fat content of the food as well as of sodium, vitamins, and other nutrients. All this has an impact on the changing eating habits of the population, as well as on the economics of the food industry.

In the health field, changes in treatment, prevention, and prevailing knowledge come about when there is a confluence of circumstances: New information is acquired to supplant existing theories; there is dissemination of this information to the scientific community and to the public at large; and there is the appropriate psychological, economic, and political climate that would welcome the adoption of the new approaches. Epidemiology plays a major role by providing the methods by which new scientific knowledge is acquired and evaluated. Often, the first clues to causality come long before a biological mechanism is known. Around 1850 in

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

London, Dr. John Snow, dismayed at the suffering and deaths caused by epidemics of cholera, carefully studied reports of such epidemics and noted that cholera was much more likely to occur in certain parts of London than in other parts. He mapped the places where cholera was rampant and where it was less so, and he noted that houses supplied with water by one company, the Southwark and Vauxhall Company, had many more cases of cholera than those supplied by another company. He also knew that the Vauxhall Company used as its source an area heavily contaminated by sewage. Snow insisted that the city closed the pump supplying the contaminated water, known as the Broad Street Pump. They did so and cholera abated. All this was 25 years before anyone isolated the cholera bacillus and long before people accepted the notion that disease could be spread by water. In modern times, the AIDS epidemic is one where the method of spread was identified before the infectious agent, the HIV virus, was known.

Epidemiologic techniques have been increasingly applied to chronic diseases, which differ from infectious diseases in that they may persist for a long time (whereas infections usually either kill quickly or are cured quickly) and also usually have multiple causes, many of which are difficult to identify. Here, also, epidemiology plays a central role in identifying risk factors, such as smoking for lung cancer and high cholesterol for heart disease. Such knowledge is translated into public action before the full biological pathways are elucidated. The action takes the form of educational campaigns, anti-smoking laws, restrictions on advertisement, and other mechanisms to limit smoking. The risk factors for heart disease have been identified through classic epidemiologic studies resulting in lifestyle changes for individuals as well as having public policy consequences. When the first edition of this book was published, the major focus of epidemiology was on chronic diseases, as it was thought they formed the major threat to health and survival and that infectious diseases were more or less under control. The world was rudely awakened from such perceptions with the advent of the COVID-19 pandemic caused by the SARS-CoV-2 virus, which spread aroud the world starting in 2020. Epidemiologic studies continue to provide knowledge about the transmission, host resistance or susceptibility, efficacy of newly discovered vaccines, and evaluation of preventative methods.

Chronic diseases present different and challenging problems in analysis, and new statistical techniques continue to be developed to accommodate such problems. New statistical techniques are also being developed for the special problems encountered in genetics research. Thus the field of statistic is not static and the field of epidemiology is not fixed. Both adapt and expand to deal with the changing health problems of our society and with advances in knowledge, as, for instance, in genetics, that require new methodologies.

Appendix 1

Critical Values of Chi-Square, Z, AND t

When Z, χ^2 , or t value calculated from the observed data is equal to or exceeds the critical value listed below, we can reject the null hypothesis at the given significance level, α (alpha).

Selected critical values of chi-square

Significance level	.1	.05	.01	.001
Critical value of χ^2	2.71	3.84	6.63	10.83

Selected critical values of Z

Significance level				
Two-tailed test	.1	.05	.01	.001
(One-tailed test)	(.05)	(.025)	(.005)	(.0005)
Critical value of Z	1.64	1.96	2.58	3.29

Selected critical values of t

Significance level				
Two-tailed test	.10	.05	.01	.001
(One-tailed test)	(.05)	(.025)	(.005)	(.0005)
Degrees of freedom				
9	1.83	2.26	3.25	4.78
19	1.73	2.09	3.86	3.88
100	1.66	1.98	2.63	3.39
1,000	1.64	1.96	2.58	3.29

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

Note: Interpretation:

If you have 19 degrees of freedom, to reject H_o, at $\boldsymbol{\alpha} = 0.05$ with a two-tailed test, you would need a value of t as large or larger than 2.09; for $\boldsymbol{\alpha} = 0.01$, a t at least as large as 3.86 would be needed. Note that when df gets very large, the critical values of t are the same as the critical values of Z. Values other than those calculated here appear in most of the texts shown in the Suggested Readings.

Appendix 2

Fisher's Exact Test

Suppose you have a two-by-two table arising from an experiment on rats that exposes one group to a particular experimental condition and the other group to a control condition, with the outcome measure of being alive after 1 week. The table looks as follows:

Moreover, 87.5% of the experimental group and 16.7% of the control group lived. A more extreme outcome, given the same row and column totals, would be

Where 100% of the experimental and 0% of the control group lived. Another more extreme outcome would be where 25% of the experimental and all of the controls lived:

(Any other tables we could construct with the same marginal totals would be less extreme than Table A.1, since no cell would contain a number less than the smallest number in Table A.1, which is 1.)

We calculate the exact probability of getting the observed outcome of the experiment by chance alone (Table A.1), or one even more extreme (as in either

Table A.1 Exposure by outcome

	Control	Experimental	
Alive	*a* 1	*b* 7	Row 1 = R_1= 8
Dead	*c* 5	*d* 1	Row 2 = R_2 = 6
Total	Col 1 = C_1 = 6	Col 2 = C_2 = 8	*N* = 14

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

Table A.2 Exposure by outcome

	Control	Experimental	
Alive	0	8	8
Dead	6	0	6
Total	6	8	14

Table A.3 Exposure by outcome

	Control	Experimental	
Alive	0	2	8
Dead	0	6	6
Total	6	8	14

Table A.2 or A.3), if it were really true that there were no differences in survival between the two groups. Fisher's exact test is calculated by getting the probability of each of these tables and summing these probabilities.

First we have to explain the symbol "!". It is called a "factorial." A number n ! means $(n) \times (n\ -1) \times (n\ -2) \times \ldots \times (1)$. For example, $6!\ = 6 \times 5 \times 4 \times 3 \times 2 \times 1$. By definition, 0! is equal to 1.

The probability of getting the observations in Table A.1 is

$$\frac{(R_1!) \times (R_2!) \times (C_1!) \times (C_2!)}{a! \times b! \times c! \times d! \times N!} = \frac{8!6!6!8!}{1!7!5!1!14!} = .015984$$

The probability of getting the observations in Table A.2 is

$$\frac{8!6!6!8!}{0!8!6!0!14!} = .000333$$

The probability of getting the observations in Table A.3 is

$$\frac{8!6!6!8!}{6!2!0!6!14!} = .009324$$

The sum of these probabilities is $0.015984 + 0.000333 + 0.009324 = 0.025641$. Thus we can say that the exact probability of obtaining the results we observed in Table A.1, or results more extreme, is 0.025641, if the null hypothesis were true. We may reject the null hypothesis that the survival rate is the same in both groups at a significance level$p = 0.026$.

Appendix 3

Kruskal–Wallis Nonparametric Test to Compare Several Groups

For example, suppose you have three groups of people each having a score on some scale. The total number of people in all three groups is N. The general procedure is as follows: (1) Combine all the scores from the three groups and order them from lowest to highest. (2) Give the rank of 1 to the lowest score, 2 to the next lowest, and so on, with N being assigned to the person with the highest score. (3) Sort the people back into their original groups, with each person having his assigned rank. (4) Sum all the ranks in each group. (5) Calculate the quantity shown below, which we call H. (6) If you have more than five cases in each group, you can look up H in a chi-square table, with k –1 degrees of freedom (where k is the number of groups being compared).

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

Scores On Reading Comprehension					
Group A		Group B		Group C	
Scores	(Rank)	Scores	(Rank)	Scores	(Rank)
98	(13)	80	(9)	120	(21)
70	(6)	60	(2)	110	(17)
68	(5)	106	(15)	90	(12)
107	(16)	50	(1)	114	(19)
115	(20)	75	(8)	105	(14)
65	(4)	74	(7)	85	(10)
(Sum of Ranks)		64	(3)	112	(18)
				87	(11)
	(64)		**(45)**		**(122)**

$$H = \left(\frac{12}{N(N+1)} \times \sum \frac{(R)^2}{n_j}\right) - 3(N+1)$$

$$H = \left(\frac{12}{21(22)}\right) * \left(\frac{64^2}{6} + \frac{45^2}{7} + \frac{122^2}{8}\right) - 3(22)$$

$$= 7.57$$

$$\text{degrees of freedom} = 3 - 1 = 2$$

If the null hypothesis of no difference in mean rank between groups was true, the probability of getting a chi-square as large as, or larger than, 7.57 with two degrees of freedom is less than 0.05, so we can reject the null hypothesis and conclude the groups differ. (When ties occur in ranking, each score is given the mean of the rank for which it is tied. If there are many ties, a correction to H may be used, as described in the book by Siegel listed in Suggested Readings.)

Appendix 4

How to Calculate a Correlation Coefficient

Individual	X	Y	X^2	Y^2	XY
A	5	7	25	49	35
B	8	4	64	16	32
C	15	8	225	64	120
D	20	10	400	100	200
E	25	14	625	196	350
Σ	73	43	1339	425	737

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5(737) - (73)(43)}{\sqrt{5(1339) - (73)^2}\sqrt{5(425) - (43)^2}} = \frac{3685 - 3139}{\sqrt{1366}\sqrt{276}}$$

$$= \frac{546}{(37)(16.6)} = \frac{546}{614} = 0.89$$

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

How to Calculate Regression Coefficients

$$b = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}; \quad a = \frac{\Sigma Y}{N} - \frac{b\Sigma X}{N}$$

$$b = \frac{737 - \frac{(73)\ (43)}{5}}{1339 - \frac{(73)^2}{5}} = \frac{737 - 628}{1339 - 1066} = \frac{109}{273} = 0.40$$

$$a = \frac{43}{5} - \frac{.40(73)}{5} = 8.60 - 5.84 = 2.76$$

Appendix 5

Age Adjustment

Consider two populations, A and B, with the following characteristics:

Population	Age	Age-specific rates	# of people in population	# of deaths in population	Crude death rate
A	Young	$\frac{4}{1{,}000} = .004$	500	$.004 \times 500 = 2$	
	Old	$\frac{16}{1{,}000} = .016$	500	$.016 \times 500 = 8$	
	Total		1,000	10	$\frac{10}{1{,}000}$
B	Young	$\frac{5}{1{,}000} = .005$	667	$.005 \times 667 = 3.335$	
	Old	$\frac{20}{1{,}000} = .020$	333	$.020 \times 333 = 6.66$	
	Total		1,000	10	$\frac{10}{1{,}000}$

Note that the population B has higher age-specific death rates in each age group than population A, but both populations have the same crude death rate of 10/1,000. The reason for this is that population A has a greater proportion of old people (50%)

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

and even though the death rate for the old people is 16/1,000 in population A compared with 20/1,000 in population B, the greater number of people in that group contributes to a greater number of total deaths.

To perform age adjustment, we must select a standard population to which we will compare both A and B. The following examples use two different standard populations as illustrations. In practice, a standard population is chosen either as the population during a particular year or as the combined A and B population. The choice of standard population does not matter. The phrase "standard population" in this context refers to a population with a particular age distribution (if we are adjusting for age) or sex distribution (if we are adjusting for sex). The age-specific (or sex-specific, if that is what is being adjusted) rates for both groups A and B are applied to the age distribution of the standard population in order to compare A and B *as if* they had the same age distribution.

Note if you use two different standard populations, you get different age-adjusted rates, but relative figures are the same; that is, the age-adjusted rates for A are lower than for B. This implies that the age-specific rates for A are lower than for B, but since the crude rates are the same, it must mean that population A is older. Because we know that age-specific rates for older people are higher than for younger people, population A must have been weighted by a larger proportion of older people (who contributed more deaths) in order to result in the same crude rate as B but in a lower age-adjusted rate.

There are exceptions to the above inference when we consider groups where infant mortality is very high. In that case it could be that the young have very high death rates, even higher than the old. In industrialized societies, however, the age-specific death rates for the old are higher than for the young.

Standard population I: Example (more old people than young)

Age	# of people		Apply age-specific death rates for population A to standard population		# of deaths expected in A if it were the same age composition as the standard population	Apply age-specific death rates for population B to standard population	# of deaths expected in B if it were the same age composition as the standard population
Young	300	×	.004	=	1.2	.005	1.5
Old	700	×	.016	=	11.2	.020	14.0
Total	1,000				12.4		15.5
Age-adjusted rates for: A = 12/1,000 B = 15/1,000							

Standard population II: Example (more young people than old)

Age	# of people	Apply age-specific death rates for population A to standard population	# of deaths expected in A if it were the same age composition as the standard population	Apply age-specific death rates for population B to standard population	# of deaths expected in B if it were the same age composition as the standard population
Young	1,167	× .004 =	4.67	.005	5.84
Old	833	× .016 =	13.33	.020	16.66
Total	2,000		18		22.50
Age-adjusted rates for: $A = \frac{18}{2{,}000} = \frac{9}{1{,}000}$			Age-adjusted rates for: $B = \frac{22.5}{2{,}000} = \frac{11.25}{1{,}000}$		

Appendix 6

Determining Appropriateness of Change Scores

(1) To determine if change scores are appropriate:

Consider a group of 16 patients who have the following scores on a scale assessing depressive symptoms; a retest is given shortly after to determine the variability of scores within patients (Table A.4):

An analysis of variance indicates the following (Table A.5):

This is greater than 0.5, so that the use of change scores is appropriate. [Note: σ^2, or the variance, is the mean square (MS) from the analysis of variance.]

Next, the patients are divided into two groups; one group is given a dietary intervention lasting 10 weeks, while the other group serves as a control group. The scale is administered again after 10 weeks to both groups, with the following results (Table A.6):

(2) To calculate *coefficient of sensitivity* to change, do a repeated measures analysis of variance on the scores in the *treatment group*; to get the error variance, calculate the variance of the change scores (Table A.7).

Coefficient of sensitivity = variance of change scores in treatment group/(variance of change scores + error variance) =

$$\frac{44.57}{(44.57 + 22.57)} = 0.66$$

(the 44.57 is obtained from in the last column of Table A.6).

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

Table A.4 Test and retest scores

Patient #	First Test Scale Score	Retest Score
1	12	13
2	16	15
3	22	21
4	24	23
5	30	29
6	18	19
7	16	15
8	12	12
9	14	15
10	18	18
11	24	24
12	30	29
13	18	19
14	16	15
15	14	15
16	10	11
Mean	18.38	18.31

Table A.5 Analysis of variance

Source of variation	SS	df	MS	F	*P*value
Patients	1014.7188	15	67.6479	156.8647	0.0000
Test retest	0.0313	1	0.0313	0.0725	0.7915
Error	6.4688	15	0.4313		
Total	1021.2188	31			

$\sigma^2_{\text{between patients}} = (67.65 - 0.43)/2 = 33.61$

$\sigma^2_{\text{between patients + error}} = 33.61/(33.61 + 0.43) = 0.987$

(3) *Effect size* = mean of the change scores/s.d. of pretest scores: in the treatment group = −11/6.32 = −1.74 (there was a *decline in depression symptom score* of 1.74 pretest standard deviation units).

(4) *Guyatt's responsiveness measure* 30 is (mean change scores in the treatment group)/(s.d. of change scores in stable subjects). We are assuming here that the control group is the group of stable subjects, although generally "stable subjects" refers to subjects who are stable with regard to some external criterion.

Table A.6 Change scores between pre- and posttest

Control Group				Treatment Group			
Patient #	Pre-test	Post-test	Change Score	Patient #	Pre-test	Post-test	Change Score
1	12	13	1	9	14	7	–7
2	16	15	–1	10	18	10	–8
3	22	20	–2	11	24	7	–17
4	24	18	–6	12	30	5	–25
5	30	25	–5	13	18	10	–8
6	18	16	–2	14	16	8	–8
7	16	12	–4	15	14	4	–10
8	12	10	–2	16	10	5	–5
Mean	18.75	16.13	–2.63		18	7	–11
Variance	38.79	23.27	5.13		40.00	5.14	44.57
s.d.	6.23	4.82	2.26		6.32	2.27	6.68

Table A.7 Analysis of variance of change scores

Source of variation	SS	df	MS	F	*P*value
Between test/retest	484	1	484.0000	21.4430	0.0004
Within patients	316	14	22.5714		
Total	800	15			

$$G = \frac{-11}{2.26} = -4.86$$

(5) *Comparison with a control group:* The effect size for the treatment group is -1.74, so clearly it exceeds the control group change, which is $-2.63/6.23 = -0.42$. If we calculate the ratios of treatment to control group for the above indices of responsiveness, we will find in this example that they are very similar.

For effect size the ratio is $-1.74/-0.42 = 4.14$. For Guyatt's statistic it is $-4.86/-1.16 = 4.19$. (The -1.16 was obtained by mean change in control group divided by standard deviation of change scores in control group, i.e., $= -2.63/2.26$.)

For the coefficient of sensitivity, it is $0.66/0.14 = 4.71$. (The 0.14 was obtained by doing an analysis of scores in the control group, not shown here, so take it on faith, or calculate it as a check on the accuracy of this.)

References

A note about thereferences:

Some of these references are quite old, for one of two reasons. Either they refer to a paper that provided data for the example being used here to illustrate a stastitical concept, or they refer to a classic article that explains the concept or method being described in a particularly clear way. More extensive general references are provided at the end of this section under the heading: Suggested Readings. Current and updated references on genetics are provided in the section on references for Chapter 8, under the heading: Additional Suggested References: Design and Interpretation of Genome-Wide Studies

Chapter 1

1. Popper KR: The Logic of Scientific Discovery. New York: Harper and Row, 1959.
2. Weed DL: On the Logic of Causal Inference. American Journal of Epidemiology, 123(6):965–979, 1985.
3. Goodman SN, Royall R: Evidence and Scientific Research. AJPH, 78(12):1568–1574, 1988.
4. Susser M: Rules of Inference in Epidemiology. In: Regulatory Toxicology and Pharmacology, Chapter 6, pp. 116–128. New York: Academic Press, 1986.
5. Brown HI: Perception, Theory and Commitment: The New Philosophy of Science. Chicago: Precedent, 1977.
6. Hill AB: and Hill ID, Bradford Hill's Principles of Medical Statistics, Hodder Education Publishers, 1991
7. Feinstein AR, Principles of Medical Statistics, New York; Chapman and Hall/CRC, 2001
   Lash TL, Vanderweele TJ, Haneuse S, Rothman KJ: Modern Epidemiology Fourth Edition; Wolters Kluwer, 2021

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2

Chapter 3

8. Drapkin A, Merskey C. Anticoagulant therapy after acute myocardial infarction. Relation of therapeutic benefit to patient's age, sex, and severity of infarction. *Journal of the American Medical Association*. 1972;222:541–548
9. Intellectual development of children. U.S. Department of Health, Education, and Welfare, Public Health Service. HSMA, Vital and Health Statistics Series 11, 1971
10. Davis BR, Blaufox MD, Hawkins CM, Langford HG, Oberman A, Swencionis C, Wassertheil-Smoller S, Wylie-Rosett J, Zimbaldi N. Trial of antihypertensive interventions and management. Design, methods, and selected baseline results. *Controlled clinical trials*. 1989;10:11–30
11. Oberman A, Wassertheil-Smoller S, Langford HG, Blaufox MD, Davis BR, Blaszkowski T, Zimbaldi N, Hawkins CM. Pharmacologic and nutritional treatment of mild hypertension: Changes in cardiovascular risk status. *Annals of internal medicine*. 1990;112:89–95
12. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1: 43–46
13. Scarr-Salapatek S. Race, social class, and iq. *Science*. 1971;174:1285–1295
14. Sokal RR RJ. *Biometry*. 4th Edition, San Francisco: W.H. Freeman; 2011
15. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: An orientation. *American journal of epidemiology*. 1971.

Chapter 4

16. Hypertension Detection and Follow-Up Program Cooperative Group, Blood pressure studies in 14 communities. A two-stage screen for hypertension. *Journal of the American Medical Association*. 1977;237:2385–2391
17. Sorlie PD, Allison MA, Aviles-Santa ML, Cai J, Daviglus ML, Howard AG, Kaplan R, Lavange LM, Raij L, Schneiderman N, Wassertheil-Smoller S, Talavera GA. Prevalence of hypertension, awareness, treatment, and control in the hispanic community health study/study of latinos. *American journal of hypertension*. 2014;27:793–800
18. Inter-Society Commission for Heart Disease Resources, Report of inter-society commission for heart disease resources. Primary prevention of the atherosclerotic diseases. *Circulation*. 1972; XLII:1–44
19. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: The Framingham study. *The American journal of cardiology*. 1976;38:46–51
20. These data come from the National Pooling Project. For purposes of this example, high blood pressure is defined as diastolic blood pressure >=105 mmHg. The disease in question is a "Coronary Event" and the time period is 10 years. Note that hypertension is currently defined as systolic blood pressure >=140 mmHg and/or diastolic blood pressure >=90 mmHg
21. Lilienfeld AM, Lilienfeld D,. Foundations of Epidemiology, New York, Oxford University Press, 1980
22. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, Kotchen JM, Ockene J. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *JAMA: the journal of the American Medical Association*. 2002;288:321–333.
23. The numerical example is courtesy of Dr. Martin Lesser, Cornell University Medical Center, New York
24. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*. 1998;17:2265–2281

25. Smoller JW, Allison M, Cochrane BB, Curb JD, Perlis RH, Robinson JG, Rosal MC, Wenger NK, Wassertheil-Smoller S. Antidepressant use and risk of incident cardiovascular morbidity and mortality among postmenopausal women in the women's health initiative study. *Archives of internal medicine*. 2009;169:2128–2139
26. Greenland S. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health.* 1989;79:340–349
27. Wassertheil-Smoller S, Fann C, Allman RM, Black HR, Camel GH, Davis B, Masaki K, Pressel S, Prineas RJ, Stamler J, Vogt TM. Relation of low body mass to death and stroke in the systolic hypertension in the elderly program. The shep cooperative research group. *Archives of internal medicine*. 2000;160:494–500
28. Eisenhauer JG. Meta-analysis and mega-analysis: A simple introduction. Teaching Statistics 2021;43:21-27. https://doi.org/ 10.1111/test.12242

Chapter 6

29. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I. Introduction and Design. *British Journal of Cancer* , 34:585–612, 1976.

Chapter 7

30. Guyatt G, Walter S, Norman G: Measuring Change Over Time: Assessing the Usefulness of Evaluative Instruments. J Chronic Dis 1987;40:171–8.

Chapter 8

31. Duncan LE, Pollastri AR, Smoller JW. Mind the gap: Why many geneticists and psychological scientists have discrepant views about gene-environment interaction (gxe) research. The American psychologist. 2014;69:249–268
32. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. The New England journal of medicine. 2000;343:78–85
33. Hinrichs AL, Larkin EK, Suarez BK. Population stratification and patterns of linkage disequilibrium. Genetic epidemiology. 2009;33 Suppl 1:S88-92
34. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet. 2006 May;7(5):385–94. doi: 10.1038/nrg1839. PMID: 16619052.
35. Uffelmann, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. Nat Rev Methods Primers 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9
36. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608. doi: 10.1002/mpr.1608. Epub 2018 Feb 27. PMID: 29484742; PMCID: PMC6001694.
37. Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, Karlsson R, Frei O, Fan CC, De Witte W, Mota NR, Mullins N, Brügger K, Lee SH, Wray NR, Skarabis N,

Huang H, Neale B, Daly MJ, Mattheisen M, Walters R, Ripke S. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. Bioinformatics. 2020 Feb 1;36(3):930–933.
38. Chen JJ, Roberson PK, Schell MJ. The false discovery rate: A key concept in large-scale genetic studies. Cancer control: journal of the Moffitt Cancer Center. 2010;17:58–62
39. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep;15(9):2759–2772. doi: 10.1038/s41596-020-0353-1. Epub 2020 Jul 24. PMID: 32709988; PMCID: PMC7612115.
40. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, Kenny EE, Pasaniuc B; Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium Methods Working Group; Witte JS, Ge T. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2023 Aug 24. doi: 10.1038/s41576-023-00637-2. Epub ahead of print. PMID: 37620596.
41. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776. doi: 10.1038/s41467-019-09718-5. PMID: 30992449; PMCID: PMC6467998.
42. Barry CS, Walker VM, Cheesman R, Davey Smith G, Morris TT, Davies NM. How to estimate heritability: a guide for genetic epidemiologists. Int J Epidemiol. 2023 Apr 19;52(2):624–632. doi: 10.1093/ije/dyac224. PMID: 36427280; PMCID: PMC10114051.
43. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson N, Daly MJ, Price AL, Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291–5. doi: 10.1038/ng.3211. Epub 2015 Feb 2. PMID: 25642630; PMCID: PMC4495769.
44. STAARpipeline: an all-in-one rare-variant tool for biobank-scale whole-genome sequencing data. Nat Methods. 2022 Dec;19(12):1532–1533. doi: 10.1038/s41592-022-01641-w. PMID: 36316564.
45. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, Hartwig FP, Kutalik Z, Holmes MV, Minelli C, Morrison JV, Pan W, Relton CL, Theodoratou E. Guidelines for performing Mendelian randomization investigations: update for summer 2023. Wellcome Open Res. 2023 Aug 4;4:186. doi: 10.12688/wellcomeopenres.15555.3. PMID: 32760811; PMCID: PMC7384151.

Additional Suggested References: Design and Interpretation of Genome-Wide Studies

Sham PC and Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15:335–46.

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S and Sunyaev S. Sequencing studies in human genetics: design and interpretation. Nat Rev Genet. 2013;14:460–70.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP and Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9:356–69.

Price AL, Zaitlen NA, Reich D and Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11:459–63.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR and Lander ES. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111:E455–64.

Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. Lancet. 2005 Sep 10-16;366(9489):941–51. doi: 10.1016/S0140-6736(05)67322-9. Erratum in: Lancet. 2006 Jan 7;367(9504):28. PMID: 16154023Friedman NP, Banich MT, Keller MC. Twin studies to GWAS: there and back again. Trends Cogn Sci. 2021 Oct;25(10):855–869. doi: 10.1016/j.tics.2021.06.007. Epub 2021 Jul 24. PMID: 34312064; PMCID: PMC8446317.

Verbanck, M., Chen, CY., Neale, B. et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet 50, 693–698 (2018). https://doi.org/10.1038/s41588-018-0099-7 PMCID: PMC6083837, NIHMSID: NIHMS947520, PMID: 29686387

Zhu X. Mendelian randomization and pleiotropy analysis. Quant Biol. 2021 Jul 13;9(2):122–132. doi: 10.1007/s40484-020-0216-3. Epub 2020 Oct 21. PMID: 34386270; PMCID: PMC8356909.

National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Health and Medicine Division; Committee on Population; Board on Health Sciences Policy; Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field.* Washington (DC): National Academies Press (US); 2023 Mar 14. PMID: 36989389.

Chapter 9

46. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: The Framingham study. *The American journal of cardiology*. 1976;38:46–51
47. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: Adjustment for antihypertensive medication. The Framingham study. *Stroke; a journal of cerebral circulation*. 1994;25:40–43
48. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *American heart journal*. 1991;121:293–298
49. Wassertheil-Smoller S, McGinn A, Allison M, Ca T, Curb D, Eaton C, Hendrix S, Kaplan R, Ko M, Martin LW, Xue X. Improvement in stroke risk prediction: Role of c-reactive protein and lipoprotein-associated phospholipase a(2) in the women's health initiative. *International journal of stroke: official journal of the International Stroke Society*. 2012
50. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clinical chemistry and laboratory medicine: CCLM / FESCC*. 2010;48:1703–1711
51. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*. 2011;30:11–21
52. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *American journal of epidemiology*. 2012;176:473–481

Suggested Reading

Cohen J: Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum, 1988[1]

Cohen J (ed), Cohen P, West SG, Aiken LS: Applied Multiple Regression: Correlation Analysis for the Behavioral Science, 3rd ed., Lawrence Erlbaum, 2002[2]

Fleiss JL, Levin B, Paik MC: Statistical Methods for Rates and Proportions, 3rd ed. New York: John Wiley and Sons, 2003[3]

Friedman, LM, Furberg CD, DeMets DL, Fundamentals of Clinical Trials, 4th edition, Springer, 2010[4]

Hosmer DW, Lemeshow S, Sturdivant RX: Applied Logistic Regression. 3rd edition, New York: John Wiley and Sons, 2013[5]

Kleinbaum DG, Kupper LL Inizam A, Muller KE,: Applied Regression Analysis and Other Multivariable Methods. Fourth edition, Thomson Books/Cole 2007[6]

Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ: Modern Epidemiology 4th Edition, Wolters Kluwer, 2021[7]

Siegel S and Castellan Jr. NJ: Nonparametric Statistics for the Behavioral Sciences. 2nd Edition. New York, Toronto, London: McGraw- Hill, Inc., 1988[8]

Sokal RR, Rohlf JF: Biometry, 4th ed. San Francisco: W.H. Freeman, 2011[9]

Streiner L, Norman GR: Health Measurement Scales. A Practical Guide to Their Development and Use, 4th edition. New York: Oxford University Press, 2008[10]

Szklo M, Nieto FJ: Epidemiology. Beyond the Basics. 3rd edition, Jones & Bartlett Learning; 2012[11]

[1] A classic and a must. Everything you ever wanted to know about sample size calculations, clearly and comprehensively explained. A reference and source book. Probably needs to be interpreted by a statistician.

[2] Excellent, thorough, and clear exposition of very complex topics. It's fairly advanced.

[3] An excellent classic, second-level statistics text concerned with the analysis of qualitative or categorical data.

[4] Classic text on the subject.

[5] Advanced book on model building in logistic regression requires statistical background.

[6] This text, as the title indicates, deals with multiple regression and allied topics.

[7] An advanced text that covers both design and statistical issues. The focus is an observational epidemiologic study and is directed at the researcher more than the clinician.

[8] A "how-to-do-it" book, an excellent reference, outlines each procedure step-by-step, a classic.

[9] Comprehensive reference work.

[10] Excellent and understandable book about scale construction and change scores.

[11] Somewhere between first and second level text, distinguished by its clear style and very readable, and many excellent examples.

Index

S. Wassertheil-Smoller, J. Smoller, *Biostatistics and Epidemiology*,
https://doi.org/10.1007/978-3-031-53043-2